The Value of Evaluation:
Tools for Budgeting and Valuing Evaluations

Julian Barr (Itad), David Rinnert (DFID), Rob Lloyd (Itad), Danielle Dunne (DFID), Anna Henttinen (DFID)

with contributions from Bob Crabtree and Tom Wagstaff

Discussion Paper

August 2016

This material has been funded by UK aid from the UK government; however the views expressed do not necessarily reflect the UK government's official policies.

itad

Department for International Development

UKaid from the British people

Executive Summary

DFID has been at the forefront of supporting the generation of evidence to meet the increasing demand for knowledge and evidence about what works in international development. Monitoring and evaluation have become established tools for donor agencies and other actors to demonstrate accountability and to learn. At the same time, the need to demonstrate the impact and value of evaluation activities has also increased. However, there is currently no systematic approach to valuing the benefits of an evaluation, whether at the individual or at the portfolio level.

This paper argues that the value proposition of evaluations for DFID is context-specific, but that it is closely linked to the use of the evaluation and the benefits conferred to stakeholders by the use of the evidence that the evaluation provides. Although it may not always be possible to quantify and monetise this value, it should always be possible to identify and articulate it.

In the simplest terms, the cost of an evaluation should be proportionate to the value that an evaluation is expected to generate. This means that it is important to be clear about the rationale, purpose and intended use of an evaluation before investing in one. To provide accountability for evaluation activity, decision makers are also interested to know whether an evaluation was 'worth it' after it has been completed. Namely, did the investment in the evaluation generate information that is in itself more valuable and useful than using the funds for another purpose.

Against this background, this paper has been commissioned by DFID to answer two main questions:

1. What different methods and approaches can be used to estimate the value of evaluations *before* commissioning decisions are taken and what tools and approaches are available to assess the value of an already concluded evaluation?

2. How can these approaches be simplified and merged into a practical framework that can be applied and further developed by evaluation commissioners to make evidence-based decisions about whether and how to evaluate before commissioning and contracting?

To answer these questions, the paper first reviews different valuation techniques from a range of academic disciplines and looks at their challenges and potential usefulness for the valuation of evaluations. Nine valuation techniques are reviewed: Value of Information analysis, Cost Benefit Analysis, Cost Effectiveness Analysis (ex-ante), Cost Effectiveness Analysis (ex-post), HERG Payback Framework, Value Chain Analysis, case studies and expert review, narrative description of benefit, and ultimate beneficiary value ranking.

Most of the valuation techniques are relatively time consuming to use, require a specific set of skills to apply, and can only be applied in a context of abundance of high quality data. While some of the techniques can generate detailed and specific estimations of value, they often do so at the expense of wider utility. This paper finds that most ex-ante techniques may be too time-consuming for evaluation commissioners, including DFID, to use routinely.

More complicated and time-consuming valuation techniques may be justified where the benefits are likely to be large, for example where the information generated by an evaluation has the potential to be massively scaled-up and used across countries and/or agencies. In contrast, some of the ex-post techniques are suitable for further adaptation and use. Drawing on this analysis, this paper presents a framework that can be further developed by evaluation commissioners into an ex-ante tool to articulate and estimate the potential benefit of evaluations that they plan to commission.

Table of Contents

1. Background and Rationale for this Paper

Many evaluation practitioners believe that "*evaluation can yield benefits well in excess of its direct costs… through wholesome effects on the productivity of expenditures and the responsiveness of the public organizations*".[1] Yet, there is surprisingly little empirical evidence available to demonstrate the benefits of evaluation, or to show they can be estimated.

A recent 'What Works Review of the Use of Evidence in the Department for International Development' notes: "*No-one will argue that using evidence is ever bad*".[2] 'Evidence' is thus usually seen as axiomatically 'a good thing'. This means that the value of evaluation can become self-evident, and the justification for evaluating unquestioning. As with other investments decision makers make, investment in evaluation requires a rationale and the spend needs to be proportionate to the expected value generated. The difficulty arises in deciding how much to spend to generate evidence; i.e. what is the value of evidence, and is this a good buy? To ensure accountability and value for money of the evaluation function and to learn lessons for future commissions, it is also important to understand if the expected value of evidence has been realised after an evaluation has concluded, and whether the investment in an evidence product has indeed been proportionate.

DFID places explicit value on use and policy relevance of its evaluations, which is reflected in DFID's Evaluation Strategy 2014-2019.[3] The Strategy defines five strategic outcomes which will deliver its evaluation goal of reducing poverty by generating evidence and knowledge that informs effective decision making. The central strategic outcome is that "*Evaluation is integral to the planning, design, and implementation of policies and programmes*". This aim of **mainstreamed use** of evaluation evidence is supported by the other four outcomes, which relate to **focus** (evaluations respond to high priority information needs), **quality** (evaluations uphold the highest quality standards)**, communication** (evaluations are actively communicated in a timely and useful way), and **capacity** (evaluation enhances partners' capacity).

The need for a strategic approach was supported by an ICAI review (2014)[4] which noted that while DFID had grown its evidence generation operations in the last few years, it was less able to explain how this evidence is used. ICAI found that DFID *"does not review the costs, benefits and impact of learning. Insufficient priority is placed on learning during implementation. DFID does not routinely assess the impact of learning on decision-making*." An ICAI observation and recommendation set a clear context and rational for this paper (Box 1).

> *"DFID needs to focus more clearly on the value of its evaluations if they are to support effectively how the organisation learns. … DFID does not track or report on the overall impact of evaluations …. We would have expected DFID also to report the impact that evaluations have on what it does and what it achieves. Such reporting would cover actions taken in response to individual evaluations and their impact on DFID's overall value for money and effectiveness."*

Box 1. ICAI (2014) Observation and Recommendation on Value and Use of Evaluations

[1] Picciotto, R. (1999) Towards an Economics of Evaluation. *Evaluation*, Vol 5 (1): 7–22.
[2] Dercon, S. (DFID Chief Economist) in: *What Works Review of the Use of Evidence in the Department for International Development (DFID)*, June 2014.
[3] DFID (2014). *DFID Evaluation Strategy 2014-2019.*
[4] ICAI (2014). *How DFID Learns*. Report 34. Independent Commission for Aid Impact, London.

DFID is not the only government department grappling with these issues. A National Audit Office (NAO) review (2013)[5] of evaluation in government was critical across its sample of departments – it found that: "*There is little systematic information from the government on how it has used the evaluation evidence that it has commissioned or produced*".

Other development agencies such as Norad[6], Australian Department of Foreign Affairs and Trade (DFAT)[7], the European Commission[8] and the United Nations Development Programme (UNDP)[9] have all recently commissioned studies that assess how to improve their evaluation systems and strengthen the uptake of evaluation findings. At the same time, many of these agencies are also aiming to understand better how to analyse the impact of their evaluations. The World Bank has been asking questions about value of evaluation[10], though they acknowledge that doing so is challenging: "*We don't have the answers. Do any of you?*"[11] There is a growing body of academic literature interested in understanding the conditions that shape evaluation use.[12] Evaluators and implementers have also started to engage with the value of evaluation agenda, outlined for example by a recent 3ie/IDinsight paper looking at how to maximise the social impact of impact evaluations.[13] More broadly, with the increase in supply of evaluations in international development, there has been an emerging interest in improving understanding of how to value the benefits and costs of evaluations, both before commissioning and after completion.

This working paper aims to help DFID and other agencies to be more explicit about the value of their evaluations. This is done through an examination of techniques that can be used in the valuation of evaluation (sections 2 and 3) and the incorporation of these techniques into a framework to guide commissioners thinking about the potential benefit of evaluations to be commissioned (section 4). It seeks to contribute to the wider discussion on the valuation of evaluation, with some of the more DFID-focussed implications that arose from this work being detailed in an internal paper

[5] NAO (2013). *Evaluation in government*. National Audit Office, London.

[6] Itad (2014) *Can we demonstrate the difference that Norwegian Aid makes? An evaluation of results measurement and how it can be improved*, Norad Evaluation Department, Report 1/2014, Oslo. http://www.norad.no/globalassets/import-2162015-80434-am/www.norad.no-ny/filarkiv/evalueringsavdelingens-filer/use-of-evaluation-in-the-norwegain-development-cooperation-system.pdf

[7] Itad (2014) *Quality of Australian Aid Operational Evaluations*, DFATD, Australia. http://www.dfat.gov.au/aid/how-we-measure-performance/ode/Documents/ode-brief-quality-aust-aid-op-evaluations.pdf

[8] ECDPM/ODI (2014) *Assesssing the Uptake of Strategic Evaluations in EU Development Cooperation*, EU, Brussels. https://ec.europa.eu/europeaid/sites/devco/files/strategic-evaluation-uptake-study-1331-main-report-201406_en.pdf

[9] Bastel (2014) *Review of the UNDP evaluation policy*. October 2014, UNDP

[10] Heider, C. (2014). *Unpacking the "Value" in the Value-for-Money of Evaluation*. IEG #WhatWorks blog. World Bank, 16/09/14. http://ieg.worldbank.org/blog/value-money-evaluation-business

[11] Heider, C. (2014). *Value-for-money in the Evaluation Business*. IEG #WhatWorks blog. World Bank, 29/07/14. http://ieg.worldbank.org/blog/value-money-evaluation-business

[12] Johnson, K., Greenseid, L.O., Toal, S.A., King, J.A., Lawrenz, F., Volkov, B., 2009. Research on Evaluation Use A Review of the Empirical Literature From 1986 to 2005. American Journal of Evaluation 30, 377–410; Ledermann, S., 2012. Exploring the Necessary Conditions for Evaluation Use in Program Change. American Journal of Evaluation 33, 159–178; Herbert, J.L., 2014. Researching Evaluation Influence A Review of the Literature. Eval Rev 38, 388–419.

[13] Shah, N. B. et al (2015). *Evaluations with impact. Decision-focused impact evaluation as a practical policymaking tool.* New Delhi: International Initiative for Impact Evaluation (3ie). http://www.3ieimpact.org/media/filer_public/2015/10/01/wp25-evaluations_with_impact.pdf

2. What do we mean by value of evaluation?

Before examining techniques to measure the value of evaluation, it is important to consider what we mean by 'value'. The concept of 'value' of evaluations is linked to **whether and how the knowledge generated during or from an evaluation will be used and by whom**. Whilst the value of evaluations can be analysed with different techniques at the levels of both the individual evaluation and the portfolio[14], the focus of this paper is the level of the individual evaluation.

2.1 Value to whom?

In any discussion of value, the question inevitably arises: "value to whom?" Recently**, Public Value** has emerged as a theory and school of public management.[15] Public Value (PV) argues for a "*renewed emphasis on the important role public managers can play in maintaining an organisation's legitimacy in the eyes of the public*". PV considers *value* as being constituted of three dimensions: 1. Delivering actual services; 2. Achieving social outcomes; 3. Maintaining trust and legitimacy of the agency.[16] PV also introduces the idea of a more democratic conception of value (i.e. public value, to the users of public services). Given that PV is an accepted way of conceiving value, we can by extension infer its dimensions of value on the value of evaluation.

This is important in the value of evaluation discussion because maintaining trust in the system is a core aspect of the accountability function of evaluation, and one that can get lost in the focus on learning and taking evidence of what works to scale. Accountability and trust is not just about trust in the use of public funds, it is also about being accountable to beneficiaries and giving beneficiaries an input into how aid programmes are delivered and how they are evaluated.

Most of the techniques to measure value discussed in this paper (section 3) tend to depend on expert views on value, although data collection in approaches such as the Payback Framework and the use of case studies could include gathering views of ultimate beneficiaries. Recognising the need to understand how beneficiaries value research impact, Professor Jonathan Grant at King's College London (KCL) is undertaking a research project on valuing research using the views of the general population. Grant argues that: "*In the absence of knowledge about - and methods for assessing - values of beneficiaries and funders of research, policy that directs researchers to estimate impact and research funders to value such impacts rests on weak empiric foundations*."[17]

---

[14] The World Bank has recently released a study analysing the effect of impact evaluations in the rate of disbursements for their projects. The main finding is that IEs increase disbursements. Legovini, A., Di Maro, V., Piza, C. (2015). *Impact Evaluation Helps Deliver Development Projects*. https://openknowledge.worldbank.org/bitstream/handle/10986/21154/WPS7157.pdf?sequence=1

[15] Moore, M. (1995) *Creating Public Value*, Cambridge, Mass., Harvard University Press

[16] Kelly, G., Mulgan, G and Muers, S. (2002). *Creating Public Value An analytical framework for public service reform*. Strategy Unit, The Cabinet Office, London.

[17] Grant, J. (2015). Understanding the relative valuations of research impact: Applying best-worst scaling experiments to survey the public & biomedical/health researcher. http://www.kcl.ac.uk/sspp/policy-institute/projects/Innovation-Policy/Value-of-Impact.aspx

2.2 Intended use of evaluation

The scale of benefit varies with intended use. Before addressing the scale of benefit, attention needs to be given to intended evaluation purpose, use and thus type of benefit. There are a number of reasons to evaluate, and several systems for categorising these.[18] The NAO summarises current Treasury advice on the role of evaluation in supporting evidence-based policy-making into three main uses[19]:

- To **inform strategic resource allocations**, such as in spending reviews;
- To **inform decisions about policies and programmes**, in terms of the design of new programmes, and improving or stopping existing programmes;
- To **support accountability**, by demonstrating the costs and benefits of spending.

An interesting schema of evaluation purposes has also been proposed by Habicht et al.[20] They suggest evaluation functions fall into categories of determining adequacy, plausibility, or probability, which essentially relate to increasingly rigorous designs. This categorisation is useful for those particularly concerned about levels of attribution and strength of causality. However, it is less useful in grouping evaluations by their primary use. The categories proposed by Habicht et al are:

- **Adequacy** - evaluations which assess whether or not the goals, set by programme developers, were met;
- **Plausibility** - evaluations which determine whether a programme has attained expected goals, yet identifies changes as potential effects of programme activities rather than external or confounding sources;
- **Probability** - evaluations which "*determine the success of a programme's activities and outcomes, but makes use of the most robust study design, randomised control trials (RCTs) to determine the true effect of the programme activities on the indicators of interest*."

Evaluations, as with any type of evidence vehicle, are used in a variety of ways in policymaking. DFID categorises use of evidence into three types, using the 'Stetler Model' of evidence-based practice[21] from the medical field. This describes three types of **evaluation use**:

- **Instrumental**: this is where the knowledge from an evaluation is used directlty to inform an ongoing policy or programme;
- **Conceptual**: this is where no direct action is taken as a result of the evaluation, but where the knowledge from the evaluation influences people's general thinking around what works;
- **Symbolic**: this occurs when people use the mere exsitence of an evaluation, rather than its specific findings, to persuade or convince. A version –

[18] There is a summary of four other schema in: Weyrauch, V and Diaz Langou, G. (2011). *Sound Expectations: from impact evaluations to policy change*. Working Paper 12. International Initiative for Impact Evaluation, New Dehli.

[19] NAO (2013). *Evaluation in government*. National Audit Office, London. http://www.nao.org.uk/wp-content/uploads/2013/12/10331-001-Evaluation-ingovernment_30-12.pdf.

[20] Habicht, JP, Victoria, CG, and Vaughan, JP (1999). Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *Int. J. Epi*.; 28, 10-18. http://www.uniteforsight.org/evaluation-course/module3

[21] Stetler, C.B. (2010). Chapter 3: *Stetler Model*. In J. Rycroft-Malone & T. Bucknall (Eds.), *Models and frameworks for implementing evidence-based practice: Linking evidence to action*. Evidence-based Practice Series. Wiley-Blackwell, Oxford.

politcal/strategic use - is when an evaluation is used to justify or legitimate a policy or decision.

Ignoring symbolic use, consideration of where evidence use is located on an instrumental-to-conceptual spectrum is important, as it has a large influence on the potential value of an evaluation.

Reviewing across a range schema for evaluation purposes and use, for this paper, we conclude that broadly, the case for evaluation rests on **three main applications** for evaluation:

- **Accountability and Value for Money**. Good governance dictates that the recipients of public funding should be able to provide an account of their decision making. The norm is that most programmes should be assessed to satisfy these requirements, and as part of good practice to demonstrate that funding has been used for the intended purposes. Within the spectrum of evaluative/assessment instruments, these requirements may be achieved by evaluation or by other means. This function comes from the Results Based Management (RBM) school and does not refer to financial audit, but rather what is often called in the UK a Value for Money (VFM) assessment or VFM audit, and elsewhere is known as a *performance audit*. It should be noted that in terms of value, evaluation may be considered as legitimating authority (government) through improved accountability[22] – it thus has a value in ensuring the right functioning of government / the public sector broadly. This should not be underestimated.

- **Internal - To improve the use of funds in a current programme**. Funding agencies and governments may require an evaluation as a condition of funding, but usually an evaluation will aim to **improve programme performance**. A number of types of evaluation are encompassed within this function, including **process evaluation, performance evaluation and others**. This internal function may also include using the evidence to close a programme or change a policy. If the evaluation shows the intervention to be ineffective, its benefit would be avoiding the costs of continuing with the programme and potentially avoiding unintended negative effects..

  In this function, the evidence generated by the evaluation will mainly be used internally *to the intervention that is evaluated*. It will be used to **improve the delivery** of the *current* intervention or policy, and thus help ensure it delivers its planned benefits within its lifetime. The benefits from the evaluation are in essence bounded by the value of the benefits the intervention is designed to create. The value of the evaluation is the net present value of the additional net benefits of the programme above what it would have delivered without an evaluation.[23] The primary use is therefore **instrumental**. This function is becoming increasingly important as development actors adopt adaptive and flexible programming approaches[24], that depend upon sound within-programme learning cycles and "decision-focused" evaluations which serve

[22] Picciotto, R. (1999) Towards an Economics of Evaluation. *Evaluation*, Vol 5(1): 7–22.
[23] If an evaluation identifies failure and recommends closure, the benefit of the evaluation is the value of potentially wasted expenditure which is not committed and can now be used for other activity.
[24] Vowles, P. (2013). *Adaptive programming*. DFID Blog, 21/10/13, https://dfid.blog.gov.uk/2013/10/21/adaptive-programming/

as "context-specific tools for decision making that feed into local solution finding systems".[25]

- **External - To determine if a policy or programme has worked**, and therefore to generate evidence and knowledge that informs decision-making more broadly, for example implementation in another context or **going to scale**. This may simplistically be considered **as learning**.

  In this function, the evidence generated by the evaluation will mainly be used *externally to the intervention*. It will **contribute learning** in an area of policy or programming, not just to improving a *particular* policy or programme. The initial use is therefore **conceptual**, and the evidence contributes (as part of a larger knowledge base) to understanding what works and why in a policy or programming area. The main types of evaluation that deliver this function are **impact evaluations**, as the majority of evidence they generate arrives ex-post, and the evidence they generate should be knowledge that can be used and more widely. The benefits can thus be amplified beyond the scope and scale of the evaluand. These evaluations create global public goods, and as Stiglitz[26] makes clear, without active public support, there will be under-provision of this good. The International Initiative for Impact Evaluation (3ie)has termed evaluations falling under this function as "knowledge-focused evaluations", which are "primarily designed to build global knowledge about development interventions and theory".[27]

Evaluations are of course not as clear-cut as these three applications suggest; most evaluations will combine an element of both internal and external use in varying proportions. Many evaluations will have an internal phase, leading to an external phase. For the purposes of the Evaluation Decision Framework, developed as part of this paper and presented in chapter four, both are considered using the **same rubrics** for scale of benefit (see below).

2.3 Benefits of evaluation

The greatest value from an evaluation is derived when the evidence it generates is used to inform a decision to increase investments in a successful policy, approach or technology (or conversely, to stop investing in a failing one). For example, in demonstrating that a particular social policy has a beneficial effect on poor peoples' lives and should be scaled up. The Zambia social cash transfer case study (see Annex 6.2.1) is an example of this; however, decision makers often depend on more than one source of evidence - policy-making is not usually as linearly related to evidence as the Zambia case would suggest.[28]

---

[25] Shah, N. B. et al (2015). Evaluations with impact. Decision-focused impact evaluation as a practical policymaking tool New Delhi: International Initiative for Impact Evaluation (3ie).

[26] Stiglitz, J. (1999). *Knowledge as a Global Public Good*. pp 308-325. In: Global public goods: international cooperation in the 21st century/ edited by Inge Kaul, Isabelle Grunberg, Marc A. Stern. UNDP, New York.

[27] Shah, N. B. et al (2015). Evaluations with impact. Decision-focused impact evaluation as a practical policymaking tool. New Delhi: International Initiative for Impact Evaluation (3ie).

[28] "*It probably takes an extraordinary concatenation of circumstances for research to influence policy decisions directly – a well-defined decision situation, a set of policy actors who have a responsibility and jurisdiction for making the decision, an issue whose resolution depends at least to some extent on information need, research that provides that information in terms that matched the circumstances within which choices will be made, research findings that are clear cut, unambiguous, firmly supported and powerful, that reach decision-makers at the time they are wrestling with the issues, that are comprehensible and*

Approaches to the valuation of research are informative for much of this paper, not least because academics have been required to demonstrate the value of their research for at least three decades. The first UK Research Assessment Exercise (RAE) was conducted in 1986, meaning debates about the value of research have developed further than the more recent ones concerning the value of evaluation.

Different types of research and evaluation studies differ in the way they consider the importance of 'use' within their valuation approaches and impact pathways. In its guidance on creating Pathways to Impact statements for academic research, Research Councils UK (RCUK) states that: "*A clearly thought through and acceptable Pathways to Impact statement should: be project-specific and not generalised; be flexible and focus on potential outcomes.*"[29] This and other parts of the guidance indicate that the expectation is for research projects to have a Pathways statement that is external and instrumental. This is essentially what the Zambia case shows. This strongly instrumental lens, wherein the findings about a particular technology or mechanism are scaled as a result of a single study or a group of studies seem uncommon, or even simplistic, and ignore the complex reality of policy making and programme design.[30] Evaluations are usually framed in terms of purpose and use, and often recognise that the resulting evidence is expected to contribute to a pool of evidence about what works.

In considering evaluation benefits, it should also be noted that DFID's Evaluation Strategy identifies a further benefit from its evaluations: the **enhanced capacity of development partners** to demand for and use evidence and to commission, design and use evaluations. This benefit should also be taken into account when making value judgements.

There is a **time dimension** to internal and external evaluation benefits. Some impacts by their nature take a long time to be revealed and measured (e.g. impacts on long-term health or some types of behaviour change). Others, such as cash transfers, have an immediate impact on welfare but may also have longer-term effects on saving and investment. Some interventions, such as health systems improvements, may likewise take a long time to reveal their impact (on health outcomes) – and so require longer periods than others for measurement. Evaluations occurring some time after an intervention has ceased operation are expensive due to greater challenges in information retrieval and interpretation. Evaluations that occur too far ex-post also face challenges of continued relevance as well as being unable to attribute impacts to the intervention to other downstream intervening variables. There is thus an ideal time at which to evaluate, that captures impact effects reliably, while still being relevant to uptake audiences. This trade-off

---

understood, and that do not run counter to strong political interests." Weiss, C.H. (1979). The Many Meanings of Research Utilization. *Public Administration Review*, 39.5: 426–31

[29] RCUK (nd). RCUK Review of Pathways to Impact. Research Councils UK. http://www.rcuk.ac.uk/innovation/impacts/

[30] Evidence informed policy making is a complex, messy and inherently political process. While evidence can play a role, it is commonly one among a number of factors at play in a policy process. The challenge that this policy-making reality creates is that assessing the value of an evaluation often means trying to value the incremental benefit of additional evaluation evidence that is used in a conceptual way to contribute to the pool of knowledge that, over time, leads to policy change or a new programming approach. The same multi-faceted context also presents challenges for attributing changes (i.e. benefit) to evaluations. Not least, evaluators and evaluation commissioners may over-subscribe change to the agency of the evaluation evidence. This can be particularly so if an overly instrumental use of the evaluation is conceptualised. The Zambia SCT evaluation is an unusual example of a very direct instrumental use of an evaluation. This problem parallels the well-rehearsed attribution/contribution problem that evaluations face in trying to assess the results produced by interventions. As the SCT case study demonstrates, even in information-rich ex-post situations, the attribution story of the evaluation ([how much] has the evaluation induced scale up?) is often not straightforward, not least as politics often 'get in the way'.

supports an argument for employing a mix of evaluation types[31], ranging from monitoring reviews to process, performance or impact evaluations“. Ideally, one might wish to split the evaluation activities over time, measuring both short-term and long-term (sustaining) benefits, though it is rare for evaluations to return several years after interventions conclude.

Where benefits from an intervention occur over several years it is important to take into account the discount rate of benefits.[32] This is an integral element in assessing the cost-benefit or cost-effectiveness of interventions (and the evaluation of interventions).

[31] DFID considers that evaluation falls into a spectrum on tools for enquiry and assessment, which ranges from audit at one end and research at the other, but encompasses monitoring, performance audit and review between these. (DFID (2013). *International Development Evaluation Policy*.

[32] DFID has separate guidance on discounting rates.

3. Valuing evaluations at different stages of the evaluation cycle

There are fundamentally different framings and techniques to establish the value proposition of evaluations depending on which point of the evaluation cycle this is done. Many, though not all, of the main valuation techniques may be employed either at planning stages (ex-ante) or after completion (ex-post).

This section explores the different techniques and the key differences of i) certainty over costs and benefits, and ii) different purposes.

3.1 Before or After

There are two main stages of an evaluation where it is useful to assess the value of an evaluation (see figure 1):

- **Planning stages/Ex-ante valuation** – the purpose of valuing evaluations ex-ante is to assess at the planning stage whether and potentially how much it is worth to invest in an evaluation or portfolio of evaluations, based on the value of the evidence it/they are likely to generate. This is therefore mainly a practical exercise to support decision-making.
- **After completion/Ex-post valuation** – the ex-post valuation of an individual evaluation can be used for making a retrospective value-based case for having conducted that evaluation, proving (or not) that it was worth the investment, thereby mainly fulfilling an accountability function for evaluations. This can also be used for communication – for example to make the case for conducting evaluations (as with the two case studies in Annex 2). However, a growing body of ex–post valuation of evaluations at the portfolio level, and their synthesis, will build an evidence base to inform evaluation planning and create a feedback loop that informs learning about commissioning more valuable evaluations.

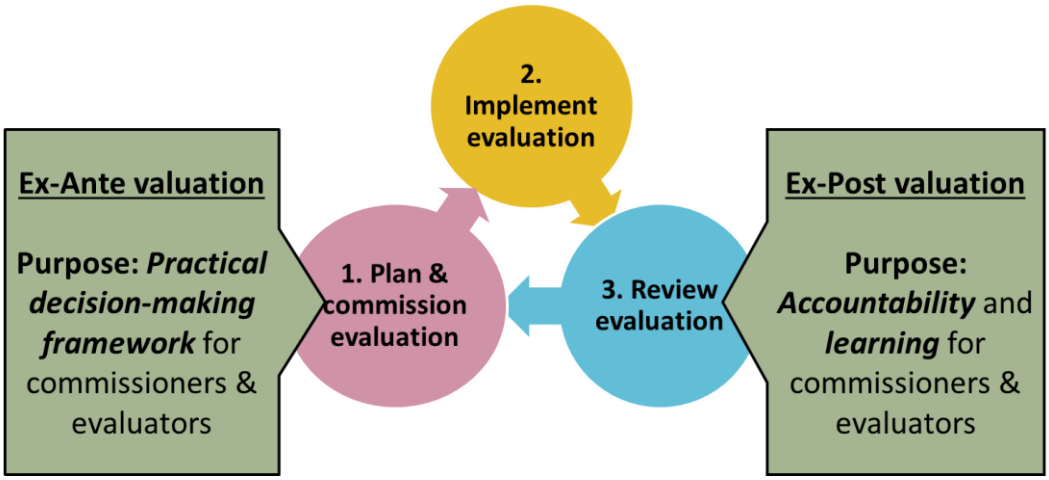


Figure 1 – The Evaluation Project Cycle and different perspectives on valuation. Source: authors of this paper.

Assessing the value of evaluation ex-ante requires commissioners to **estimate the likely use and impact** , and to try to value those. Ex-post, it requires retrospectively trying to attribute value to the evidence generated by evaluation. Although in many cases, ex-post valuation will also have a predictive component since ex-post assessments are often conducted too soon after an intervention for all the impacts to have been realised.

Ex-ante valuation compares *budgeted* costs of the investment in evaluation with its likely *expected* benefits.[33] This assessment can help determine the most appropriate design for the evaluation and how much to invest – Section 5 of this paper addresses this area. Ex-post assessments support making a judgement on whether the investment in evaluation was in fact worthwhile and delivered the expected benefits against actual costs. Both types of valuations are likely to be useful, for both DFID and the wider evaluation community.

3.2 Valuation techniques

Drawing on economic appraisal guidance and VFM work[34], benefits may be assessed in three ways. They can be **described qualitatively**, **quantified** and/or **monetised**. Different assessment **tools** can be applied depending on the information available and the feasibility of quantifying and expressing benefits in monetary terms. These are shown in Table 1, and outlined in the proceding sections.

	Ex-ante	Ex post
Monetised	• Value of Information analysis • Cost-benefit analysis (CBA)	• Value Chain Analysis • Ultimate beneficiary value ranking • Cost-benefit analysis (CBA)
Quantified	• Value of Information analysis • Cost-effectiveness analysis (CEA)	• HERG Payback Framework analysis • Value Chain Analysis • Ultimate beneficiary value ranking • Cost-effectiveness analysis (CEA)
Described qualitatively	• Narrative on potential or realised benefit	• HERG Payback Framework analysis • Questionnaire & standard indicator tool • Case study & expert review • Narrative on potential or realised benefit

Table 1. Techniques for assessing evaluation benefits

Of these approaches, some have been developed to be particularly used in an ex-post context, others ex-ante, and some may be used in either perspective:

[33] Annex 6.4 offers guidance on what information needs to be collected to support an ex-post assessment.
[34] E.g. DFID Smart Rules, and Julian Barr and Angela Christie (2014) *Better Value for Money. An Organising Framework for Management and Measurement of VFM Indicators*, Hove: Itad, http://itad.com/knowledge-products/better-value-for-money/

- **Ex-ante**: Value of Information (VoI) analysis was specifically developed to predict the value of information yet to be collected.
- **Ex-post**: The HERG Payback Framework, Value Chain Analysis, the use of questionnaires and indicators, and of case studies with expert review, are all approaches that have been developed to make ex-post assessment of the value of research. It may be possible to use all of these in a predictive manner, asking respondents to use the tools to estimate benefits yet to be created rather than report benefits already achieved, but these methodological adaptations have not been tested, and are likely to be overly data hungry.
- **Either:** Cost Benefit Analysis (CBA), Cost Effectiveness Analysis (CEA) and narrative descriptions of benefits may each be used ex-ante or ex-post, the difference being the confidence that can be ascribed to the monetisation, quantification or qualification of benefits.

There are some common challenges in ex-ante valuation approaches. These are listed below, rather than repeated for each ex-ante approach:

i) **Limited information:** there is inevitably less information available ex-ante. No evidence or benefits have been generated – these need to be predicted. No costs have been incurred, so budgets need to be estimated. Tools are needed to support making informed estimates.

ii) **Tools trade-off:** Some very sophisticated ex-ante valuation tools exist, such as Value of Information analysis. Theses require high skill levels in economics and Bayesian statistics to use properly. In general, if predictive tools are too complex or too expensive, they will not be used, But if they are too simple, they do not support good decision making. The ex-ante tools need to be sufficiently fine grained to be useful, but not so intensive as to not be adopted.

iii) **Benefits are 'potential'**. Ex-ante valuation techniques predict the potential benefit an evaluation will generate. This assumes a good evaluation design is conducted professionally, reported in a clear and robust way, and communicated in a timely way to an audience with capacity and resources to use the evidence. Many factors can thus interfere to prevent realisation of the potential evaluation benefits. These internal and external factors need to be well managed.

The following sections outline the context in which these approaches have been developed and used, and offer a short commentary on their suitability for valuing evaluations.

3.2.1 Cost-Benefit Analysis (including risk assessment) - Ex-Ante

[Ex-Ante, Monetised]

CBA is a '*policy assessment method that quantifies in monetary terms the value of all the consequences of a policy [or project] to all members of society*'. Its primary function is to help social decision making, specifically by facilitating more efficient use of society's resources. Two major types of CBA are ex-ante CBA - the standard use of the technique to support decisions about allocation of scarce resources, and

ex-post CBA - which contributes to learning about whether particular types of investment are worthwhile.[35]

Investment in evaluation can be estimated **ex-ante** against a counterfactual of no evaluation – business as usual. Evaluation is an investment decision in which costs are incurred with the aim of producing information that results in a marginal increase in the stream of future benefits which more than justify the cost.[36] Evaluation investment takes place in a context of probability and risk as the outcome of the evaluation or how it will be used are not generally known with certainty. Hence evaluation investment can be analysed as (i) a type of insurance against future investment in unsuccessful interventions (loss avoidance), and (ii) an information generating activity that increases the expected net benefit from the current programme and/or future interventions ('learning'). The economic benefit of evaluation is therefore restricted to improving the expected net benefit from subsequent phases of a programme and/or extension of similar intervention mechanisms in the future.

The net benefit of an evaluation in the context of a single scale-up/extension can be represented as

$-C + p_1$ (mean net present value loss) $+p_2 (1-p_1)$ (mean increase in net present value) where:

C = present cost of evaluation

p_1= ex-ante probability of loss (NPV<0)

p_2= ex-ante probability of increasing the net present value (NPV) of the project

These probabilities would be determined by the prior information about expected programme performance and the experience with the learning function of evaluations. This formula can be extended to include the probability of increasing the more general information base although it would be difficult to attach a value to this impact.

Taking the example of a cash transfer programme with minimal innovation and risk we might assign a low probability of failure and a low probability of learning. In this case the break-even evaluation cost would be low. For an innovative project p_1 and p_2 would be higher which would justify a higher investment in evaluation. Evaluation for very risky projects can be justified more easily[37].

Utility of this approach

This approach, whilst conceptually informative, is difficult to apply because the probabilities may be difficult to estimate objectively. In addition, the NPVs of many interventions cannot be estimated with precision because it may be difficult to express the benefits in monetary terms. We conclude that using CBA is not a suitable method for valuing most evaluations ex-ante.

[35] *Ibid.*

[36] In conventional theory only prospective (future) costs are relevant to an investment decision and it is irrational to consider sunk costs. Even so, in practice it may be that sunk costs still require ex-post justification.

[37] This is different from justifying the programme itself even though this could still be possible.

3.2.2 Cost-Benefit Analysis – Ex-Post

[Ex-Post, Monetised]

When expenditure on evaluation is assessed **ex-post**, the methodology used is similar to the ex-ante approach. In the ex-post analysis, there is the implicit assumption that only one outcome (the observed outcome) was possible. This is a major assumption for projects with uncertain outcomes[38]. The question then arises as to what the counterfactual for the assessment should be. The most obvious choice is a 'no evaluation' baseline. However, this is problematic because without an evaluation the outcomes of the single intervention (although largely unquantified and perhaps unknown) would be similar as if an evaluation had taken place.[39] There are a number of possible alternatives:

- The evaluation may have resulted in programme modifications during programme implementation. The cost of the evaluation can then be compared with benefits delivered from the adjustments.
- The evaluation may have resulted in an ineffective programme being stopped, in which case the cost of the evaluation can then be compared with the money saved from the earlier closure of the programme.
- The evaluation may have informed a scaling up or extension phase which would not have taken place without the evaluation evidence. This scaling up would typically include improvements in programme performance resulting from the evaluation. Where the evaluation measures the benefit from the initial intervention the anticipated net benefit from the next phase(s) can be compared with the cost of the evalution in a CBA:

The net benefit of the evaluation in the context of a scale-up/extension is –C + NPV where [40]

C = present cost of evaluation
NPV = net present value of the scale-up/extension[41]

If the net present value of the scale-up extension exceeds the cost of the evaluation one might conclude that the evaluation was worthwhile. Alternatively, if the cost of the evaluation exceeds the net present value of the scale-up, the evaluation investment would appear to have been unsatisfactory according to this assessment. This highlights the weakness of attempting to assess evaluation expenditure ex-post. The evaluation cost is a sunk cost and so long as the scale-up has an NPV>0 it should proceed regardless of the size of the investment in evaluation. Box 2 below outlines an example of an ex-post CBA of a Social Cash Transfer evaluation (see Annex 2 for more details on this case study).

[38] In the extreme example of insuring buildings against fire, ex-post analysis would indicate that the payment should not have been made in the great majority of cases.
[39] There has recently been some discussion in the evaluation literature about the performance effects on an intervention of merely undertaking an evaluation, even before there have been any findings reported (see Legovini, A., Di Maro, V., Piza, C. (2015). *Impact Evaluation Helps Deliver Development Projects*). While the World Bank paper shows that there is a process effect of having an evaluation on timeliness of spending, there is currently no robust evidence of a process impact of evaluations on outcome indicators,
[40] The NPVs are the present values of the discounted future benefit stream, and so embody discounting
[41] A CBA could also be used to assess improvements to a programme resulting from an evaluation.

As an example of quantifiable returns to evaluation, DFID funded an impact evaluation to measure the impacts of a social cash transfer pilot in Zambia. The programme provides a monthly £7 cash transfer to extremely poor and vulnerable people. The evaluation demonstrated a strong impact on poverty, food security, material wellbeing of children, and crop and livestock production. Following the publication of the evaluation, the Government of Zambia announced an 800% increase in its funding to its social cash transfers programme. This will allow the programme to expand from 61,000 to 190,000 recipients, benefiting approximately 950,000 people. Assuming that in the absence of the evaluation, the government would have continued the SCT programme without scaling up, and would have used remaining funds on programmes in which 80% of the spending would have reached the poorest, the return to the evaluation turns out to be between $15m – $36m

This analysis looked at benefits of the social cash transfer to the poorest households in Zambia. The findings are based on a simplified cost-benefit analysis (CBA) of the social protection programme itself, based on distributional weights (discount rate: 10%, time frame: 2013-2020, distributional weight varying between 1.5 and 3; NPV for distributional weight of 2: USD 25,167 million). This CBA tries to estimate the subjective welfare increase of the beneficiary households, net of any programme costs. Distributional weights are based on the intuition that an additional pound of income is worth more to someone on a low income than on a high income. The value that the distributional weight should take can be calculated by comparing the income of recipients with the income of taxpayers – the larger the difference, a greater distributional weight can be justified. The attribution of the SCT scale up to the evaluation results has been assessed qualitatively and more details can be found in the detailed case study.

Box 3. Zambia Social Cash Transfer example. See Annex 2 for more details.

Challenges with this approach

There are three key challenges that emerge from the two ex-post case studies undertaken for this report (Annex 2).

i) **Attribution:** Attributing changes in policy or programme design to an individual evaluation ex-post is difficult and only possible through qualitative narrative analysis. As in the social cash transfer example, there will always be a host of factors contributing to social impact, and the quantitative analysis cannot account for these fine-grained nuances. In the Zambia Social Cash Transfer (SCT) case, there were a number of political economy factors (a new government; removal of fuel subsidies; and the wider policy framework for poverty reduction and reduction of agricultural subsidy), and evaluation management factors (good timing – policy window; tightly focused and relevant study; engaged users). As such, only a qualitative judgement could be made about how much the evaluation contributed to the policy scale up.

ii) **Counterfactual:** Calculating the NPV of an evaluation, one has to make assumptions around *a counterfactual scenario*, which in many cases will be arbitrary. Rarely will it be clear what would have happened in absence of the evaluation, both in terms of costs and benefits.

iii) **Measurement:** Measuring the benefits stemming from an evaluation will be difficult in many sectors, such as in the SCT example. Beyond issues around quantification and how far down the results chain one should look for and claim benefits from an evaluation – this talks to the value chain analysis approach – is the commissioner only concerned with impact, economic and social

improvements in poor people's lives, or should outputs and outcomes be monetised also? Another question is about benefits and costs *to whom* should be taken into account. In the SCT case study, a decision had to be taken on what benefits to measure, and who was benefitting. This is particularly problematic for cash transfers as, at the basic level, there is no net benefit, just a transfer of funds. Therefore to measure benefits, an assumption of distributional weights that benefit the poorest had to be made (see annex for details).

However, as demonstrated in the second case study in Annex 6.2, the extent to which these three challenges are actually relevant when calculating NPVs of evaluations ex-post differs, depending on the type of evaluation. For example, if evaluations analyse similarly effective interventions with differences only in cost-efficiency most of the measurement and counterfactual challenges can be overcome and, more conveniently, cost savings can be calculated.

Utility of this approach

Ex-post CBA can be helpful to calculate a *range* of NPVs that can give a rough estimate of the proportionality of an evaluation, and can help to demonstrate the impact of an evaluation.

3.2.3 Cost-effectiveness analysis (CEA)

[Ex-Ante and Ex-Post, Quantified]

Cost-effectiveness analysis examines relative costs and benefits for delivering a specified programme or project. It is less demanding than CBA because it does not require the benefits to be monetised. Costs need to be monetised but this is generally straightforward. In principle it can be applied (ex-ante or ex-post) to investment in evaluation by comparing alternative designs (and costs) which are intended to deliver the same evaluation information.

Challenges with this approach

The challenge with this approach is that it is unusual for different evaluation designs to deliver the same information and data. In an ex-post approach the question 'was an evaluation cost-effective?' might be posed. CEA can be applied to answer this if there were other evaluation designs that might have been used to deliver the same information. But usually it requires an approach that takes benefits into account (as discussed above). The malaria bed net evaluation case study in Annex 6.2 looks at potential ex-post CEA of evaluations in the health sector.

Utility of this approach

Similarly to CBA, CEA requires considerable assumptions about the likely effect, scale and uptake of the evidence if used for valuing evaluations before commissioning because of the limited information and high uncertainty ex-ante. Ex-post CEA (see annex 6.2) can be helpful to demonstrate the value an evaluation has had, and its implementation is usually less demanding than ex-post CBA.

3.2.4 Value of Information analysis

[Ex-Ante, Monetised and Quantified]

Value of Information (VoI) analysis is concerned with estimating the gain that would be generated from gathering additional data to reduce the uncertainty surrounding a decision. As such, it is an ex-ante assessment approach only. VoI is used in a range of disciplines from environmental risk analysis to geo-engineering and has recently been applied to inform research priorities in the field of health care technologies[42].

VoI analysis describes the cost of uncertainty in decision making in terms of the probability that a wrong decision is made and the associated consequences of this. The Expected Value of Information (EVI) for a decision has to exceed the cost of collecting additional information in order to make additional investigation worthwhile. Further data collection is suggested if it reduces the likelihood of making the wrong decision.

To perform a VOI analysis the following information is needed:

- The full range of decision options and the information that is needed to inform these decisions.
- The likely consequence of each decision option: for each option the expected consequence needs to be defined. This needs to be quantified using monetary or non- monetary metrics.
- An estimation of the level of uncertainty that surrounds a decision.

Challenges with this approach

VoI is growing in application across a number of fields, but it is not without its challenges. Not least among these is the complexity of the methodology[43], which requires a high level of technical skills in the application of Bayesian statistical analysis.[44] In addition, conducting a full VoI analysis is a demanding exercise requiring significant investments of time and money.[45] In some instances, the costs of the VoI can exceed the costs of the research under consideration, particularly when looking at lower cost research efforts, such as systematic reviews.[46] This complexity, some argue, undermines its practicality as a day-to-day tool for decision makers.

Utility of this approach

Using VoI methodology to decide whether to invest in an evaluation or not, is an interesting approach, but the implementation of the full methodology seems **overly**

[42] A group of academics in the Centre for Health Economics (CHE) at York University (UK) specialise in this, with a focus on valuing decisions in areas of interest to the National Institute for Clinical Excellence (NICE) E.g. Karl P. Claxton, K.P. and Sculper, M.J. (2006). Using Value of Information Analysis to Prioritise Health Research: Some Lessons from Recent UK Experience. *PharmacoEconomics*, 24 (11), 1055-1068.
Zaid Chalabi, Z., Epstein, D., McKenna, C. and Claxton, K. (2008). Uncertainty and value of information when allocating resources within and between healthcare programmes. *European Journal of Operational Research*, 191 (2), 530-539.

[43] Yokota and Thompson describe the complexity of the methodology in the following way: "The application of method is very complex: to undertake the calculation one must "model all relevant sets of actions and information collection strategies available to the DM, capture all significant consequences of each action given all possible states of the world, value those outcomes in a common metric, and characterize important uncertainty, variability, and the accuracy of information to be collected by fitting probability distributions to available information."

[44] Yokota, F. and Thompson, K.M (2004) Value of Information analysis in environmental risk management decisions: past, present and future, *Risk Analysis*, Vol 24, No 3.

[45] Hoomas et al. (2012*) Systematising the use of value of information analysis in prioritizing systematic reviews*, Methods Research Report, Agency For Healthcare Research and Quality, US.
Claxton, K., Neumann, P.J., Araki, S.S. and Weinstein. M.C. (2000). *Bayesian Value-of-Information Analysis: An Application to a Policy Model of Alzheimer's Disease*. Discussion Papers in Economics No. 2000/39, Department of Economics and Related Studies University of York.

[46] *ibid*

complex for the task at hand. However, there are efforts, such as the one currently being led by the Evidence for Policy and Practice Information (EPPI) Coordinating Centre, to adapt and simplify the VoI methodology.

3.2.5 Health Economics Research Group (HERG) Payback Framework

[Ex-Post, Quantified and/or qualified]

The Payback Framework was originally developed by Martin Buxton and Stephen Hanney at the Health Economics Research Group (HERG) at Brunel University, UK, to examine the 'impact' or 'payback' of health services research.[47] To date, it appears to have been mainly used in an ex-post manner, though the framework could be used to structure prospective valuation exercises.

The Payback Framework consists of two elements:a logic model representation of the complete research processes (for the purposes of research impact evaluation) – this is essentially a theory of change for evidence uptake, and a series of categories to classify the individual paybacks from research[48]. The logic model traces the research from an initial research idea, through the research process to dissemination and thence on towards its impact on society - potentially reaching eventual final health and economic benefits. In this regard, it is essentially a variaton on the Value Chain Analysis approach outlined below. Data are collected, using a range of techniques, on effects at different parts of the logic model (Figure 1), for example on:

- knowledge production (e.g. peer-reviewed papers);
- research capacity building (e.g. career development);
- policy or product development (e.g. input into official guidelines or protocols);
- sector benefits (e.g. impacts on specific client groups); and
- wider societal benefits (e.g. economic benefits from increased population health or productivity).

The idea is that value is created in various forms at different points along the (non-linear) logic model (Figure 2). The framework is thus a research tool to facilitate data collection (by informing surveys, interview schedules and documentary analysis) and cross-case analysis by providing a common structure for each case study, thereby ensuring cognate information for each study is recorded in the same place[49]. Results on particular research (or evaluation) studies can then be compared, for example using spider diagrams (Figure 3).

[47] Donovan, C. (2016). http://www.brunel.ac.uk/herg/research-programme2/evaluating-the-payback-or-impact-from-expenditure-on-research
[48] Donovan, C. and Hanney, S. (2011). The 'Payback Framework' explained. *Research Evaluation*, 20 (3), 181–183.
[49] *ibid*

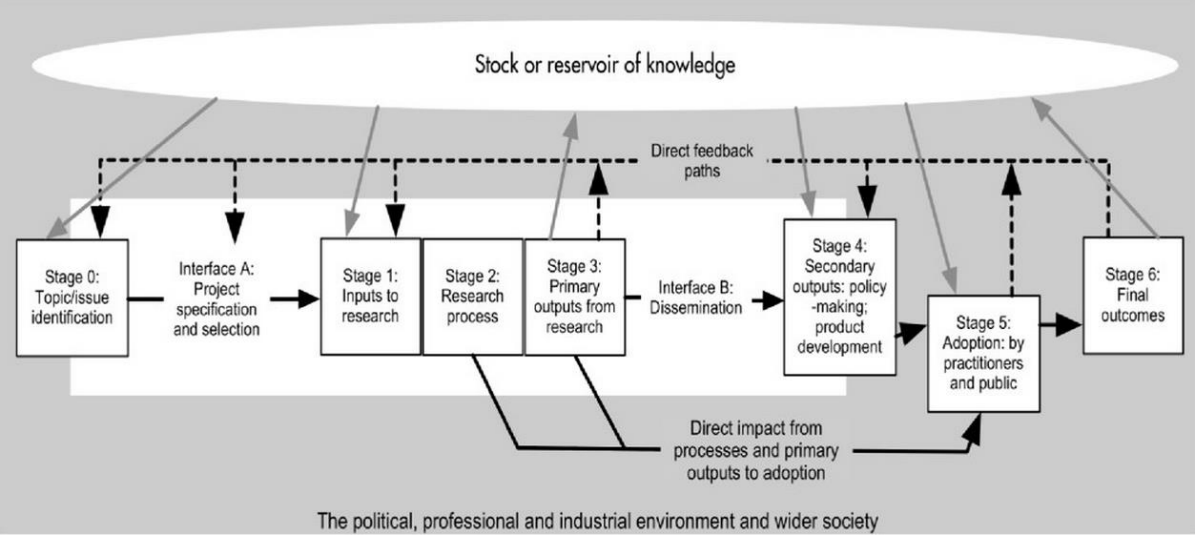
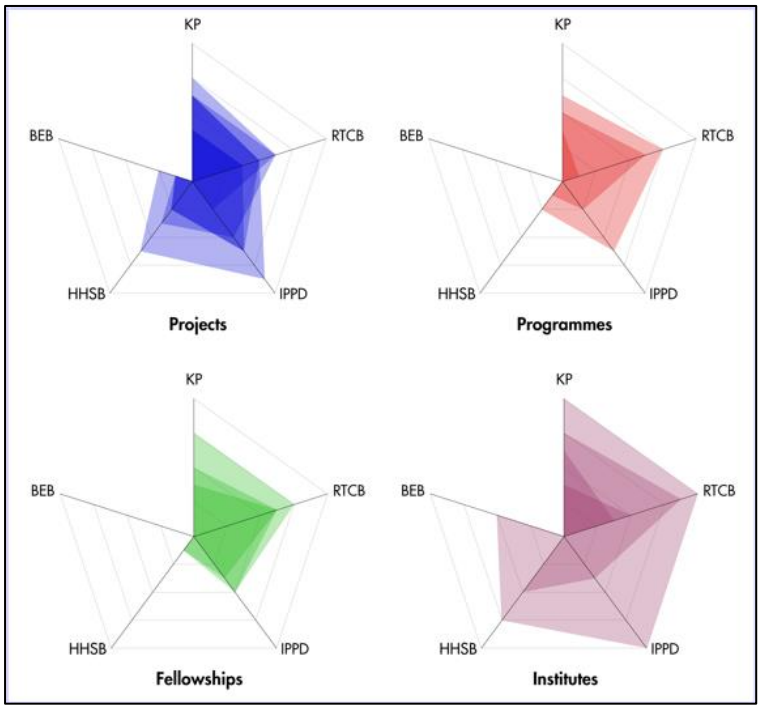


Figure 2. The logic model of the HERG Payback Framework



Key:	KP Knowledge Production RTCB Research Targeting and Capacity Building IPPD Informing Policy and Product Development HHSB Health and Health Sector Benefits BEB Broader Economic Benefit

Figure 3. Spider plot showing 'payback profiles' according to four types of funding [Source: Wooding et al, 2005[50]]

The axes in Figure 3 relate to the stages of the Payback Framework logic model: knowledge production (KP) through health and health sector benefits (HHSB). These diagrams are like-for-like comparisons of two different areas of health research.

[50] Wooding S, Hanney S, Buxton M and Grant J (2005) Payback arising from research funding: evaluation of the Arthritis Research Campaign. *Rheumatology,* 44:1145-1156.

The most frequently used data collection methods for the Payback Framework are qualitative: documentary analysis and in-depth interviews. Interviewing can be useful for understanding many aspects of research utilisation, including tracing networks between researchers and users. Some studies use insider knowledge, but there has also been some adoption of questionnaires to researchers about the utilisation of their work[51]. Using this approach, the benefit (or 'payback') of the research does not appear to be comprehensively monetised. Benefits are of a range of types, depending on the component of the logic model (e.g.: knowledge – number of publications; capacity building – PhDs supervised; health benefits - Quality Adjusted Life Years (QALYs); economic benefit – reduction in days off work due to illness [not monetised]).

The Payback Framework has been trialled outside the health sector, as a tool in assessing the policy and practice impacts of the ESRC 'Future of Work' research programme.[52] The trial concluded that: "*with minor modification, the Payback Framework can be applied to evaluate the wider impacts of social science*".[53] The logic model element of the Payback Framework, used to capture the research narrative, could be effectively applied without modification. However, unsurprisingly given their health background, the payback categories required some generalisation to fully capture the impacts of social science research. The modifications used in the Framework are shown in Table 2:

Payback category	Definition
Knowledge	Explicit and codified knowledge; papers, books and books chapters as proxy indicators
Impacts on future research	Generation of new research questions; development of new methods and/or datasets; capacity building; career development
Impacts on policy	Effects of research on policy at many levels, e.g.: national policy; policy of organisations and professional bodies – includes effects on the ability and propensity of policy makers to use research
Impacts on practice	Effects on individual behaviour, which may or may not be in line with the policies of the organisation to which the individual belongs
Wider social and economic benefits	Social or economic effects that change society, including impacts on public opinion. Media coverage can be used as a proxy indicator of public opinion.

Table 2. Revised payback categories for social science

Challenges with this approach

The drawback with the approach is that data collection is largely qualitative (although quantitative tools can be used), and this makes it time consuming to use and more difficult to compare across a range of evidence products. Also, it does only offer limited guidance on how to actually place values on benefits using the collected data, and how to compare these to potential costs.

[51] Hanney, S.R., Gonzalez-Block, M.A., Buxton, M.J. and Kogan, M. (2003). The utilisation of health research in policy-making: concepts, examples and methods of assessment. *Health Research Policy and Systems* 2003, **1**:2. http://www.health-policy-systems.com/content/1/1/2

[52] Wooding, S., Nason, S., Klautzer, S. Rubin, J, Hanney, S, Grant, J. (2007). *Policy and practice impacts of research funded by the Economic and Social Research Council A case study of the Future of Work programme, approach and analysis*. Technical report by the RAND Corporation, RAND Europe, Cambridge.

[53] Ibidem

Utility of this approach

The HERG Payback Framework provides **a standardising model for comparing benefit across evaluations**, however in its current form, it is **a research tool**.

3.2.6 Value Chain Analysis

[Ex-Post, Monetised and Quantified]

As outlined in the HERG Payback Framework, different types of benefits to research (and evaluation) accrue along the length of the research value chain. Valuing research does not consider merely end-user economic benefits to society (although some studies have done exactly that[54]). Economic, social and environmental benefits are all considered, as are other direct and indirect benefits, along the value chain - these values are often quantified rather than monetised. This approach has been used extensively in assessing the value of Australia's national investment in scientific research[55]. The Commonwealth Scientific and Industrial Research Organisation (CSIRO) considers that its value lies in:

- 'The flow of delivered research outcomes and research based advisory services
- the building and maintenance of potentially valuable research capabilities (skills, research infrastructure, networks, databases and other collections)
- the systems and internal cultures that allow these capabilities to be managed to add value to Australia's innovation efforts'

CSIRO's approach has not been to try to assess these benefits systematically across all its activities, rather, it has identified a sample of **case studies** in which these benefits have been described and quantified. A more recent study[56] has adjusted the assessment method to include, in the case studies, a CBA of direct benefits and description and quantification of indirect benefits. Challenges here are similar to those outlined above for CBAs generally.

It is worth noting that total value chain analyses are also used in the private sector to establish the holistic impact of businesses. PwC has developed the Total Impact Measurement and Management (TIMM) tool to help business understand their social, environmental, tax, and economic impacts in a holistic way[57]. To do this, they use 'leading validation techniques' to monetise businesses' impact across 20 dimensions within these four categories (Figure 4). Being a proprietary approach, PwC are not explicit about their 'leading valuation techniques'.

Utility of this approach

[54] Brennan, J. and Quade, K. (2004). *Analysis of the Impact of CIMMYT Research on the Australian Wheat Industry*. Economic Research Report No. 25. NSW Department of Primary Industries.
[55] ACIL Tasman (2010). *Assessment of CSIRO Impact & Value*. Report for CSIRO. ACIL Tasman Pty Ltd. Melbourne.
[56] ACIL Allen (2014). *CSIRO's Impact and Value. An Independent Evaluation*. A Report to CSIRO. ACIL Allen Consulting PTY LTD, Brisbane.
[57] PwC (2013). *Measuring and managing total impact – strengthening business decisions for business leaders*. PwC, London. http://www.pwc.com/gx/en/sustainability/publications/total-impact-measurement-management/assets/pwc-timm-for-ceos.pdf

In essence, this approach is an attempt at an encyclopaedic, whole-of-business ex-post or end-of-year CBA. While comprehensive, this technique is **too resource intensive** to use routinely in attempting to value evaluations.

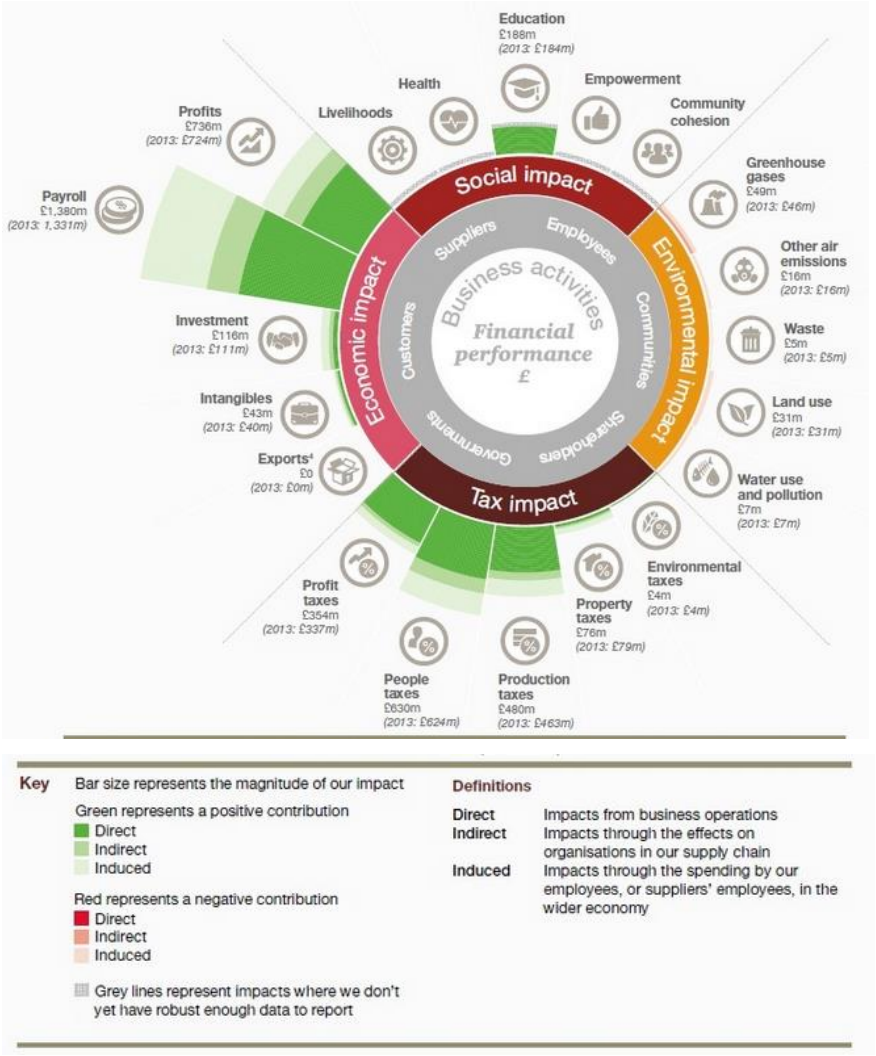


Figure 4. PwC TIMM, on PwC impact in 2014.

3.2.7 Case studies and expert review, plus questionnaires

[Ex-Post, Quantified and Described qualitatively]

Globally, in academia and the research field more widely, there has been a move away from bibliometrics (see Box 3) as key indicators of research excellence, towards case study based approaches (as with CSIRO, above) and use of qualitative approaches.[58]

In recognising the weaknesses of bibliometrics, attempts have been made to develop new metrics, such as 'technometrics' - that relate to the early stages of commercialisation of research or technology transfer; these include metrics such as: number of start-up companies created, number of patents created with industry. Technometrics favour science, technology, engineering and

[58] Grant, J. (2006). *Measuring the benefits from research*. RAND Policy Resource, RAND Europe, Cambridge.

medicine (STEM) research in terms of its value to business, industry and the economy (Donovan, 2007[59]). An analogous concern faces development evaluation, in that many existing metrics and approaches, such as CBA, favour disciplines with better established monetisation and effectiveness quantification techniques, such as health, education and infrastructure, rather than fields such as governance.

Box 3. Bibliometrics

RAND (Grant et al, 2009) reviewed four new approaches to research assessment in support of the development of Higher Education Funding Council for England's (HEFCE) new Research Excellence Framework (REF). Earlier methods, in the Research Assessment Exercise (RAE) relied strongly on bibliometrics, citation impact factors, and introduced the concept of esteem factors. The new methods included:

- The proposed Australian Research Quality and Accessibility Framework (RQF) – a case study based approach.
- The RAND/ Arthritis Research Campaign Impact Scoring System (RAISS) from the UK, which takes the form of a questionnaire (to be filled in by researchers) to capture over 150 potential research impacts.
- The widely used US Federal Program Assessment Rating Tool (PART). This questionnaire asks programmes to assess themselves against their own strategic (impact) goals.
- The Dutch 'Evaluating Research in Context (ERiC)' approach, which combines several evaluation approaches: self-evaluation; an indicator-based approach; and stakeholder analysis.

RAND concluded that Australian RQF model was the most promising for informing the REF. Thus, while the REF does still consider the impact factors of publications, it has a strong case study-based element; each department must submit a certain number of 'Impact Case Studies'. The impact case studies do not need to be representative of the spread of research activity in the unit rather they should provide the strongest examples of impact[60].

The REF Impact Case Study submissions must follow a standard template, which includes a requirement for: *"a narrative, with supporting evidence, to explain:*

- *how the research underpinned (made a distinct and material contribution to) the impact*
- *the nature and extent of the impact*."

It also requires a list of sufficient sources that could, if audited, corroborate key claims made about the impact of the research. In this context, REF defines impact as "*including, but is not limited to, an effect on, change or benefit to:*

- *the activity, attitude, awareness, behaviour, capacity, opportunity, performance, policy, practice, process or understanding*
- *of an audience, beneficiary, community, constituency, organisation or individuals*

[59] Donovan, C. (2007). The qualitative future of research evaluation. *Science and Public Policy*, 34(8), pages 585–597
[60] REF (2012). *Assessment framework and guidance on submissions*. REF2014, Research Excellence Framework, Bristol.

- *in any geographic location whether locally, regionally, nationally or internationally.*"[61]

Utility of this approach

Case study-based approaches have probably been among the most widely used methods to valuing evaluations ex-post. Their data requirements and effectiveness differ from approach to approach. However, in particular when using expert-based methods it is important to minimise biases resulting from subjective perceptions about the nature and extent of the impact of an evaluation.

3.2.8 Narrative description of benefit

[Ex-Post, Described qualitatively]

A number of types of development intervention have found it difficult to develop monetised and quantified measures of benefit. Examples may include sub-national governance projects, where, although there are national indices of quality of governance[62], these do not exist sub-nationally. Therefore, they have used qualitative descriptions of change – using tools such as **rubric scales**[63] to describe in narrative form what good governance looks like, on a scale of 1 – 5. These are qualitative measures of benefit – in this case of improved quality of governance – but using rich narrative text in rubrics to scale the benefits. Other approaches have taken truly narrative directions, with some attempting to use software, such as SenseMaker, to organise and analyse the stories[64]. The advantages of these story-based narrative approaches is that they can be more participatory with beneficiaries and accommodate plural views[65].

Utility of this approach

Interventions that *only* employ narrative descriptions of benefit as their success measures are uncommon. While programmes may have some narrative / qualitative indicators, they will normally also have some quantitative ones, and even indicators that may be monetised. Best practice, at least ex-post, is to tell a hybrid story with narrative and numbers – this is essentially what the case studies technique (3.2.7 does).

3.2.9 Valuation by ultimate beneficiaries

[Ex-Post, Monetised and Quantitative]

[61] There is an interactive database of the impact case studies submitted to REF 2014: http://impact.ref.ac.uk/CaseStudies/Results.aspx?UoA=24#.

[62] See the World Governance Indicators: http://info.worldbank.org/governance/wgi/index.aspx#home

[63] See Better Evaluation for more details: http://betterevaluation.org/evaluation-options/rubrics

[64] Kirkland, K. (2012). How alternative storytelling can help impact evaluation. *The Guardian*, 30.10.12. http://www.theguardian.com/voluntary-sector-network/2012/oct/30/storytelling-impact-evaluation

[65] Field, J. (n.d.) *Evaluation through storytelling*. The Higher Education Academy. http://jisctechdis.ac.uk/assets/documents/resources/database/id473_valuation_through_storytelling.pdf
Charles McClintock, (2004). Using Narrative Methods to Link Program Evaluation and Organization Development. *The Evaluation Exchange*, Volume IX, Number 4, Winter 2003/2004. http://www.hfrp.org/evaluation/the-evaluation-exchange/issue-archive/reflecting-on-the-past-and-future-of-evaluation/using-narrative-methods-to-link-program-evaluation-and-organization-development

All the other techniques discussed in this section tend to depend on expert or economists' views on value, although data collection in approaches such as the Payback Framework and the case studies could include gathering views of ultimate beneficiaries. This set of techniques interacts directly with the beneficiaries or users to ascribe value to evaluations is from the family of Stated Preference Methods (SPM) in economics.

To value non-market impacts of policies, particularly their value to society more widely, the Green Book[66] recommends use of Social Cost Benefit Analysis (SCBA) techniques, specifically Stated Preference Methods (SPMs) and Reveal Preference Methods (RPMs)[67]. RPMs uncover estimates of the value of non-market goods by using evidence of how people behave in the face of real market – this is not seen as applicable to evaluation. Methods within the SPM family do have applicability, particularly Contingent Valuation, which ask beneficiaries or users about willingness to pay for a service.

There is a growing body of experience of using contingent valuation-based methods within development evaluation, notably through the use of Social Return on Investment (SROI) techniques[68,69] . To date, these have been used for valuing development benefits[70], rather than the benefit of evaluation per se.

Utility of this approach

There are challenges in applying this to international development. Unlike UK health research, it is more difficult to ask the views of the UK public as they are neither the ultimate beneficiaries of development programmes, nor able to directly scrutinise the work[71]. This therefore requires specific interaction with beneficiaries in developing countries, but asking them to rank the benefits of evaluation may still present challenges, compared to asking them to assess the value of programmes per se. However, cost-benefit approaches may employ SCBA, including Social Return on Investment, to generate the benefit side of CBA calculation in an hybrid approach.

3.3 Summary

Whichever technique is used to estimate the value of benefits from evaluation, the calculation method should be **proportionate**. More complicated and time-consuming valuation techniques may be justified where the benefits are likely to be large, for example where the information generated by an evaluation has the potential to be massively scaled-up and employed across countries and/or agencies. Most of the tools described above are relatively time consuming to use, and a number require a specific set of skills to apply. Moreover, they can only be applied in a fairly limited set of data-rich situations. While they may help to generate detailed and specific estimations of value, they do so at the expense of wider utility. Taking into account

[66] HM Treasury (2011). *The Green Book. Appraisal and Evaluation in Central Government*. The Stationary Office, London.
[67] Fujiwara, D. and Campbell, R. (2011). *Valuation Techniques for Social Cost-Benefit Analysis: Stated Preference, Revealed Preference and Subjective Well-Being Approaches A Discussion of the Current Issues*. HM Treasury and Dept. for Work and Pensions, London.
[68] The SROI Network (2012). *A guide to Social Return on Investment*.
[69] Context international cooperation (2010). *Social Return On Investment. A practical guide for the development cooperation sector.* Utrecht, Netherlands.
[70] NEF consulting limited (2013). *Filling the Gaps: A social return on investment analysis*. NEF, London
[71] IFS (Fitzsimons, Emla; Rogger, Daniel and Stoye, George) (2012) 'UK Development Aid', in Carl Emmerson, Paul. Johnson and Helen Miller (eds), *The IFS Green Budget:February 2012*, London: Institute of Fiscal Studies: Chapter 7

these factors, a rubric-based approach is used in Section 3 of the Evaluation Decision Framework as the **ex-ante technique** for assessing the potential benefit of evaluations to be commissioned.

There is some **useful learning** arising from the examination of the above techniques (especially the REF impact case study approach), which could be useful to evaluation commissioners, and in particular DFID:

- Consider **selecting a sample of evaluations for ex-post valuation** within any given reporting period.
- Qualitative approaches that **include questionnaires and self-evaluation** may offer some merits for commissioners in setting up guidance to standardise the way **ongoing and ex-post information** is collected on evaluations for ex-post assessment of the benefits of evaluations.
- Consider using a **case study template** for valuing DFID **evaluations**.
- An **ex-ante valuation framework** is included in this paper (see section 4) which incorporates information from the examination of the above techniques and recommendations. Commissioners could use this framework to develop a tool, to assess the potential benefit of evaluations to be commissioned.

4. The value of evaluations and decisions on whether, why and how to evaluate

This section builds on the assessment of techniques for valuation of evaluation and simplifies and merges these into a framework which covers key considerations for deciding whether to evaluate or not, and how. The framework:

- Presents three main uses of evaluation: accountability, internal learning, and external learning;
- Structures thinking around the benefits (value) that an evaluation can potentially deliver. This is based around the scale of benefits and the probability of uptake of evidence;
- Identifies internal and external factors that need to be managed or addressed to ensure that the *potential* benefits of an evaluation are realised.

4.1 The Evaluation Decision Framework[72]

There are four stages in the framework:

1. **Justification**: Examines whether an evaluation is justified. Is there a sound rationale for evaluating, based on the importance of the intervention and the need for the evidence the evaluation will generate?
2. **Evaluability**: Is the evaluation feasible?
3. **Benefit and cost**: How will the evidence from the evaluation be used, what is the scale of the benefit and the likelihood that the evidence will be used? What is the likely level of cost, and is this proportionate given the anticipated value generated?
4. **Realising value:** How well have various quality and uptake factors been addressed that could erode or augment the projected benefits from the evaluation?

[72] The EDF is intended to promote discussion and potentially encourage commissioners to further develop this thinking into a tool , and as such does not represent DFID Evaluation Policy. Rather, this working paper puts forward suggestions based on the literature review and analysis.

The Evaluation Decision Framework

	#	Criteria	Questions
***Justification*: Strategic case for evaluating**	**1**	Strategic importance: spend & policy priority for spending unit or the commissioning organisation	• Does the programme make a significant contribution to the results set out in the spending unit's Operational Plan? • Is this a strategic and/or high profile policy commitment which needs to be evaluated? • Is there a significant investment of financial resources in this programme or policy? • Has the intervention been identified as a priority for the commissioning organisation, other than for the reasons above?
	2	Uncertainty: Evidence, Risk and Innovation	• Does the programme address an evidence gap related to an identified area of policy or programming? • Are programme outcomes and impact uncertain due to the level of risk, requiring a more comprehensive investigation of results? • Is the evaluand particularly innovative, so that there is a pressing need for evidence about its efficacy?
Evaluability	**3**	Evaluability & feasibility	• Is the proposed evaluand evaluable[73] • Is it feasible to conduct an evaluation? This mainly relates to likely access to data and information and to security issues.
***Potential Value*: Benefits and costs of evaluation**	**4**	Accountability	• Does this evaluation have an accountability purpose? • Will it provide good evidence for **VFM** and the **efficient** and **effective** use of public funds? • Will it support maintenance of **public trust** in the use of development funds?
	5	Learning & improvement: • Scale of benefit	• **How many people** are likely to benefit from the decision the evaluation aims to benefit? • What is the **level of investment** resources within the commissioning organisation that use of the evaluation evidence could affect? (refer to criteria #1 of the EDF) • How relevant is the evaluation to the policies and practices of **other external and government agencies**?
	6	Learning & improvement: • Probability of use of evidence	• Have the **eventual users** of the evaluation evidence been identified and when and how they will make use of the evidence? Is the evidence likely to be timely? • To what extent is there **an expressed demand** for the evaluation findings from the eventual users? To what extent can sound **evaluation uptake pathway be constructed**: communications & influencing strategy, political economy analysis, range of evaluation products? • How strong are the **relationships** between the evaluators, the evaluation commissioners and the eventual users
	7	Cost & Proportionality	• What is your top-down **estimate of the evaluation cost**? • Are the costs **proportionate** to the likely benefit?
***Realising* the benefits**	**8**	Quality	• To what extent are strong **quality management processes** in place: evaluation commissioner oversight, good evaluation governance structure, steering/advisory panel, QA system?

[73] This question can be answered e.g. through an evaluability assessment, see: Davies, R. (2013). *Planning Evaluability Assessments. A synthesis of literature with recommendations*. Working Paper 39. Evaluation Dept, DFID, for more details. An evaluabilty assessment refers to the assessment or feasibility of an evaluation. Evaluability assessments provide a way to determine whether an evaluation is appropriate, due to the environment, the design or other factors. The assessment can help clarify the programme logic and coherence, clarify data availability and help to identify issues for the conduct of an evaluation, including whether the stakeholders and managers are likely to use the results.

The criteria around benefits and costs and the realisation of these benefits **bring a new focus on the value of evaluations**. The following sections consider the criteria in the EDF in more detail, considering both the literature, drawing on section 3 of this paper and new and existing evaluation decision support approaches.

4.2.1 Justification – EDF criteria #1-2

The questions in this part of the EDF help determine whether it is justified to undertake an evaluation. In part, it relates to the features of the intervention (normally a programme or policy), and whether this is important, expensive, high profile, risky, evidence-based, and/or innovative, both in relation to the spending unit's and the commissioning organisation's priorities. In relation to the evaluation, it considers whether the evidence generated is likely to fill an identified gap in the knowledge. It also introduces the idea that *on its own* an evaluation may be less justified, but as part of a *cluster of evaluations*, it may make important evidence contributions on a particular theme or type of interventions, which are additively important.

4.2.2 Evaluability – EDF Criterion #3

Only where the intervention is determined as being evaluable is it worth commissioning an evaluation. In some cases, an evaluability assessment will be part of a contracted evaluator's inception phase, but more usually, evaluability will be assessed prior to commissioning. A related step here is to determine whether an evaluation is feasible[74]. DFID has recently commissioned a synthesis of the literature on evaluability assessments, which provides more detailed guidance on which questions to ask in such an assessment[75]. If a programme is deemed to be unevaluable the commissioning process should not proceed.

4.2.3 Assessing Benefits– EDF Criteria #4 - 7

This part of the EDF considers the **three main purposes** for evaluation:

- **Accountability *:*** To demonstrate that funding has been used for the purposes intended.
- **Internal learning and improvement*:*** To improve the use of funds in a current programme. There are a number of forms of evaluation that are encompassed within this, including process and performance evaluation This purpose may also include using the evidence to close a programme or change a policy. If the evaluation shows the intervention to be ineffective, its benefit would be in avoiding the costs of continuing with the programme or doing harm. The evidence generated by the evaluation will mainly be used internally *to the intervention*. It will be used to improve the delivery of the *current* programme, and thus help ensure it delivers (or exceeds) its planned benefits within its lifetime.

[74] The main considerations here are normally access to data and information and security concerns. If data access is addressed in the evaluability assessment, as it may be, then the consideration is of security and duty of care.

[75] Davies, R. (2013). *Planning Evaluability Assessments. A Synthesis of the Literature with Recommendations.* Working Paper 39. DFID, London and East Kilbride.

- **External learning and improvement*:*** To determine if a policy or programme has worked, and therefore to generate evidence that informs decision-making, particularly about going to scale. The evidence generated by the evaluation can be used *externally to the intervention*. It will **contribute learning** in an area of policy or programming, not just to improving the particular policy or programme evaluated. The primary use is therefore **conceptual**, and the evidence contributes to understanding what works, for whom and how in a policy or programming area.

Accountability - EDF criterion #4

Evaluation commissioners should consider if the evaluation is likely to:

- Provide good evidence for efficient and effectice use of public funds
- Provide evidence on the Value for Money
- Provide evidence on the performance of the intervention against its planned targets
- Support maintenance of public trust in the use of development funds
- Whether there is scope for accountability to partners, beneficiaries and host governments

Learning and improvement - EDF criteria #5-6

In the EDF, the learning and improvement function is translated into **two practical steps**; first, assessing the scale of potential benefits from internal and external learning coming from an evaluation and second, assessing the likelihood that evaluation evidence will be taken up and thus deliver the potential benefits. To examine the potential **scale of the effects** that are likely to be generated from the use of the evaluation, three proxy indicators have been developed:

1. Number of people that are likely to directly benefit from the decision that the evaluation aims to influence
2. The level of current and potential future investment resources within commissioning organisation's that the use of the evaluation could affect
3. The extent to which the evaluation is relevant to external agencies outside of commissioning organisations.

This is purposely a **group** of complementary but not overlapping indicators; together they **consider the size of the market for the evidence** that the evaluation will generate. The first is a measure of developmental benefit; a quasi-impact indicator – number of people whose lives are improved. Considerations should also be given to the magnitude of the benefit, on which information should usually be available in the business case. The second measures how important this area is for the commissioning organisation's investment; a quasi-monetary indicator – how much resource does the commissioning organisation have earmarked in this area? The third is intended to reflect that many potential users of the evaluation evidence may work in agencies other than commissioning organisation – a quasi-replication/amplification indicator.

The second step is to explore the **probability that the evaluation will be used**. guided by three considerations:

1. The clarity that exists around **who** is going to use the findings, **how** they will use them, and **when** this use is going to happen
2. The extent to which there is a clear **demand for evaluation findings** among the eventual users
3. The strength of the relationships between the evaluators, the evaluation comissioners and the eventual users – **the uptake pathway**, which will be different for internal / external audiences

The aim of this cluster of indicators is to **consider the likelihood that the evidence will be used (strength of the evidence pathway)**. The first indicator examines how well sighted the evaluation is on the users of the evidence – are target individual users already known, or are users generically 'policy makers and other stakeholders'? The second indicator examines the extent to which the proposed evaluation is supply-driven, or whether a genuine demand for the evidence it will generate has already been expressed. Thirdly, the final indicator examines the strength of engagement of the users with the design, conduct, and presentation of the evaluation, as well as in its oversight – engaged users are known to be more likely to utilise the evidence.

An evaluation that is likely to have both a significant scale of effects and high probability of uptake has *very high potential value*. It is likely to have relevance to a wide audience, both inside and outside the commissioning organisation and has the potential to influence significant investment resources; it is focused on an issue that is a major policy area; and a large number of people are likely to benefit from the decisions that use the evaluation. In addition, the conditions that make the use of the evaluation more likely are in place, therefore the probability that the evaluation is used is also high.

If an evaluation is likely to have a low scale of effects and low probability of uptake it is likely to have *very low potential value*. Both the scale of the effects and the probability of use are low which makes the investment not only risky, but also of limited benefit. It may be an interesting problem to understand, but the learning is of limited value. Commissioning organisations should avoid commissioning this type of evaluation.

4.2.4 Realising the Benefits – EDF criterion #8

Consideration should be given to how to maximise the likelihood that the benefits will be realised: "*What can be done to move from a situation where evaluation reports sit on shelves gathering dust – or worse; they are misused – how can we move from this to a situation where evaluations contribute to "social betterment"?*[76]

Evaluations may be conceived as having a high potential benefit if the evaluation results are likely to benefit a large number of people; the evaluation products are timely and appropriately tailored to the intended audience etc. However, evaluation

[76] Simon Hearn (2013) *How can evaluation make a difference?* BetterEvaluation portal, 05/07/13. http://betterevaluation.org/blog/report-support-use

commissioners often express concerns about the extent to which these ideals are in fact attained in many evaluations[77,78]. The potential benefit of evaluations may be eroded or enhanced through a range of factors and these factors may be thought of as '**modifiers**' of an evaluation's value. Modifiers can be considered as either within the control of the evaluation comissioner (internal) our outside of their control (external). While the former can be actively managed, the latter need to at least be considered and strategies developed for mitigating the potential risk they pose.

Key **internal** modifers include:

- **Quality of the evaluation** – the intrinsic quality of an evaluation is a key modifier of its potential value. Evaluation quality factors include: clarity about the evaluation purpose and objectives, the selection of a team with the right mix of skills and expertise, the appropriateness and 'right rigour'[79] of the methodology, the robustness of the data and the analytical frameworks, ensuring possibility of verifiable findings, the extent to which plural views are represented in the evaluation findings[80], and the use of the evaluation findings to draw useful conclusions and recommendations.
- **Clarity and appropriateness of how the evaluation evidence will be communicated to users –** unless the commissioner is clear on who the eventual users of the evaluation are and how they will use the evaluation, the value of the process is likely to be eroded. However, even if the users are clearly identified, it is important to ensure the evaluation is communicated with a range of evaluation products. Evaluation communication needs to be actively managed, and is likely to include a combination of appropriate evaluation products tailored to the audience **throughout the evaluation cycle** such as: in-house and conference presentations, meetings with users, blogs, intranet articles, peer-reviewed publications, summary reports, etc. It is well established that evaluation evidence is more likely to be taken up if users have been engaged throughout the evaluation process[81]. An evaluation that is delivered to an uninterested and unresponsive audience, or is rebutted, is unlikely to have traction and therefore unlikely to generate benefit.
- **Timing of the evaluation –** being clear on the timing of key decisions that the evaluation is to influence is central to ensuring it delivers value. Evaluation evidence needs to be made available prior to the decision-making processes it is intended to inform; late delivery of evaluations which have an instrumental use is of limited benefit.

Key **external** modifiers include:

- **The capacity of users to engage with and use the evaluation –** the ability of decision makers to engage with the evaluation and act on its findings is an

[77] Marlène Läubli Loud and John Mayne (2013). *Enhancing Evaluation Use. Insights from Internal Evaluation Units*. Sage Publications, Thousand Oaks and London.

[78] Ole Winckler Andersen (2014). Some Thoughts on Development Evaluation Processes. *IDS Bulletin*, Vol 45 (6), 77-84.

[79] Clemens, M. (2012*). Impact Evaluation in Aid: What for, how rigorous?* Presentation at Royal African Society and Overseas Development Institute, July 3, 2012

[80] This, and other of these modifiers, are considered in the factors used by SEQAS to assess evaluation quality.

[81] Michael Quinn Patton (2008). *Utilization-Focused Evaluation: 4th edition*. Sage Publications, Thousand Oaks

important external modifier of an evaluation's potential benefit. This is frequently identified as a key barrier to evidence informed policy making more generally.[82]

- **The political economy of evidence –** Evidence-based policy making, or even evidence-informed policy making (EIPM[83]), is easily over-simplified into an evidence supply problem: supply of more and better evidence is incontrovertibly a *good thing*, and will therefore be used to improve the quality of policies. This overlooks the many factors that are affected by the political economy of policy-making. An evaluation may deliver a high quality evidence product, but it may be side-lined because it does not resonate with past experience of decision makers or the findings may run counter to a key policy position of the government. Failing to understand the political economy of the decision / policy making process that an evaluation is intended to influence can severely degrade the potential benefits it can generate.

Summary

The Evaluation Decision Framework (EDF) presented here incorporates findings from the review of techniques (in section 4). It is expected that this framework will promote discussion in the evaluation community and potentially lead to development of a simple ex-ante tool that can support value-based decision making in evaluation commissioning

[82] Itad (2015) *How can capacity development promote evidence-informed policy making?*: Evidence review for the Building Capacity for Research Use Programme, DFID, London, UK. [Draft]

[83] Kirsty Newman, Antonio Capillo, Akin Famurewa, Chandrika Nath and Willie Siyanbola (2013). *What is the evidence on evidence-informed policy making?* Lessons from the International Conference on Evidence-Informed Policy Making. INASP, Oxford.

5. Conclusions and Recommendations

This paper was commissioned to help DFID address a demand for being able to better articulate the value of the evaluations it commissions. It examined nine different valuation techniques[84] and fed in to a framework that covers key criteria to consider when examining the value of evaluations. Overall, this paper concludes that:

- **Evaluation has lagged behind** research in making a value-based case for itself.
- **Approaches are needed** to support a *value-of-evaluation* lens on evaluations.
- It should always be possible to identify and articulate the value of an evaluation, and in some cases it will also be feasible to quantify or monetise it.
- **There are some promising valuation techniques**. Most of the techniques reviewed are relatively time consuming to use, and a number require a specific set of skills to apply. Moreover, they can only be applied in a fairly limited set of data-rich situations. While they may help to generate detailed and specific estimations of value, they often do so at the expense of wider utility. This is particularly true of the ex-ante techniques, such as Value of Information analysis, which are likely to be too involved for many evaluation commissioners. However, some of the ex-post techniques are promising and would be suitable for commissioning organisations, such as DFID, to use. CBA has been tested in two case studies (see Annex 6.2), and case study based approaches - like the Research Excellence Framework - could be very useful, particularly if a portfolio of cases if built over time.
- Ex-post CBAs of evaluations are feasible, but not straightforward. The CBA needs to sit within a wider understanding of context and the evaluation value chain. It is proposed that DFID undertakes a sample of **ex-post valuation case studies** each year.
- Whichever technique is used to estimate the value of benefits from evaluation, the **calculation method should be proportionate**. More complicated and time-consuming valuation techniques may be justified where the benefits are likely to be large, for example where the information generated by an evaluation has the potential to be massively scaled-up and employed across countries and/or agencies.

Three recommendations emerge from this paper. They may be relevant to other commissioning organisations, but are made specifically with DFID in mind:

Particular valuation techniques:

1. Many of the valuation techniques reviewed (in Section 3) are not suited to DFID's needs. However, a number of these are being adapted for development evaluation, or could be piloted by Evaluation Department (EvD). DFID should follow and review cases where particular valuation techniques are being used in

[84] Value of Information analysis, Cost Benefit Analysis, Cost Effectiveness Analysis (ex-ante), Cost Effectiveness Analysis (ex-post), HERG Payback Framework, Value Chain Analysis, Case studies and expert review, Narrative description of benefit, and Ultimate beneficiary value ranking.

an evaluation context, and assess their relevance for incorporating into DFID's valuation toolbox. For example: EPPI's attempt to adapt and simplify the VoI methodology, and the work at King's College London on beneficiary contingent valuation of impact case studies.

For ex-post valuation:

2. Ex-post valuation offers a great potential to learn about the value of evaluation. DFID should build a rolling portfolio of ex-post valuation studies, with the two case studies in Annex 6.2 of this report as the first of these.
3. The 2014 academic Research Excellence Framework exercise is an example of the excellent resource that can be built with a portfolio of impact studies. It is suggested that DFID samples a certain number of evaluations that complete each year and undertakes an ex-post case study, including a CBA on these. EvD could develop a common case study template and the case studies could be peer-reviewed within the cadre. Despite having a common reporting template, it would be useful to try different valuation techniques in these case studies.

   The evaluation of REF's / HEFCE's introduction of impact case studies usefully considered the relative balance of "the benefits and the burden" of assessing the impact of academic research (as described earlier, using case studies). This is a useful lens for DFID. There would be cost implications for DFID were it to introduce a routine exercise of ex-post case studies / CBAs of completed evaluations. Nonetheless, there is indeed a strong argument for having better sight of the ex-post value of evaluations. Such an exercise would have considerable benefit to DFID, and DFID should consider annually sampling its population of completed evaluations to conduct such ex-post valuation studies..

   As the portfolio of ex-post studies grows, DFID should develop an overall map and other schematic presentations to summarise the case studies. These are powerful communication tools. Examples of such maps conducted by CSIRO has on its impact case studies[85] (Figure 7), Jonathan Grant, on the 2014 REF impact case studies (Figures 8 and 9) are in Annex 6.5. In this paper, he analysed the 6,975 impact case studies submitted to REF 2014[86]. It is informative to see the type of graphical summary that can be produced from this very rich set of case studies. This is another example of what DFID could do if, over time, it builds a portfolio of ex-post valuations of evaluations.[87]

[85] Branwen Morgan (2014). Research impact: Income for outcome. *Nature,* 511, S72–S75. (24 July 2014)

[86] Grant, J., & Hinrichs, S. (2015). *The nature, scale and beneficiaries of research impact: An initial analysis of the Research Excellence Framework (REF) 2014 impact case studies*. Research Report 2015/01 King's College London and Digital Science, Prepared for HEFCE.

[87] In early 2016, DFID EvD has initiated a work stream on case examples of the use of evaluations.

6 Annexes

6.1 Terms of Reference

Terms of Reference

Introduction

DFID seeks a contractor to undertake a study to identify the most appropriate methods and tools for identifying and valuing the costs and benefits of evaluations.

Costing evaluations has been a concern for DFID and other donors for decades, but with increasing use of evaluation designs that involve large-scale primary research (e.g. RCTs) and growing interest in evidence from development programmes there is increasing demand for guidelines on how to budget for evaluations. At the same time, it is even more difficult to value or quantify the benefits of an evaluation, and to calculate its overall net present value (the present value of the differences between the streams of costs and benefits of evaluations). Against this background, the existing literature is of limited use as it does not provide answers to most of the emerging questions (see below).

The main purpose of this study is to provide advice to DFID, other donors and policy makers on how to budget for different types of evaluations, and beyond this, on how to identify and calculate the NPV of evaluations. The study feeds into DFID's new evaluation strategy 2014-2019, in which DFID has committed itself to "apply a more consistent approach to the decision to evaluate programmes while recognising local stakeholder priorities".

Audience

The audience for the report are DFID advisers and decision makers, global policy makers and other stakeholders involved in commissioning and managing evaluations in development and beyond. Study findings and recommendations need to be presented in a way that is accessible for both technical and non-technical audiences.

Objectives and Scope

The contractor is expected to deliver a short report (maximum of 25 pages excluding annexes) that provides answers to the following questions:

A) Costing:

1) What are the cost components of evaluations and what are the current average standard unit costs, charging practices and expected number of expert working days of these components in international development, particularly (but not exclusively) for:
    - a) Primary research (e.g. costs of surveys, data management, etc.)
    - b) Consultancy activities for secondary research and evaluation
2) What are approximate unit cost differences between different types of evaluations and between different regions/countries DFID is working in?
3) Building on the answers to the above-mentioned questions, which methods and tools can and should be used to budget for different types of evaluations for international development programmes (e.g. develop a budgeting template)?

B) Benefits/valuation:

1) What are the potential benefits of evaluation (e.g. policy influence, decision-making for intervention design, ongoing improvements in delivery, cross-project learning, post-project evidence, etc.)? And what is the incremental cost-

effectiveness of evaluating a programme, which would e.g. come from the benefit of the decision that is taken on the basis of the evaluation process or findings (i.e. value of the information that an evaluation provides)?

2) What are the benefits of the evaluation process, i.e. "process use". What benefits can derive from evaluation as an intervention in its own right?
3) How can the benefits of evaluations be valued and a NPV be calculated (e.g. expected NPV of the stream of net benefits arising from the policy or programme decisions taken in light of the information arising from the evaluation minus expected NPV of the stream of net benefits arising from the policy decisions that would have been taken at that time in the absence of the evaluation)? And how can a counterfactual be identified against which the benefits are being tested?
4) Building on the answers to the above-mentioned questions, identify a minimum of 3 DFID sample programmes together with DFID advisers and outline how identification and valuation of benefits of evaluations work in practice. How can the results from these case studies inform a better understanding of value for money of evaluations in general?

The contractor should draw on existing literature and practices of donors, including benchmarking where possible. The report should include recommendations and tools that can be used by DFID advisers and other donors for budgeting for and the valuation of future evaluations. The contractor is expected to work together closely with a DFID evaluation adviser that will provide the necessary data and information within a reasonable scope. At the same time, DFID expects the contractor to provide annexes with a short analysis of sensitive data that will not be for public consumption.

Deliverables and outputs

The following deliverables and outputs are expected as part of the project:

- An inception report/analytical framework for the report. This must include - **To be completed within 2 weeks of contract signing**:
  - An overview on the existing literature and the data/information to be used for compiling the study;
  - Clear methodology for answering the questions set out above, and, where necessary, refinement of these questions
  - Overview of the analytical framework and method(s); identification of a minimum of 3 sample programmes for the study in discussion with DFID adviser(s)
  - Work-plan and timeline for completing the study;
  - Dissemination/communication plan for the final report;
  - *Given the focus of the study, the consultant is expected to be in close touch with DFID during inception period (e.g. at least two calls) to avoid different interpretations of the scope.*
- Draft Final report – **To be completed within 8 weeks of contract signing**;
- Final report, incorporating suggestions and revisions to the draft final report – **To be completed within 10 weeks of contract signing**;
- Presentation of the report to DFID and/or external audiences and participation in any pre-agreed dissemination/communication events. - **Dates to be agreed with supplier and DFID Evaluation Department.**

Methods

The analysis and conclusions contained in the report should be based on the following:

- Desk review of academic (economics/evaluation) papers on budgeting for/valuation of research and evaluations, evaluation and programme documents from DFID and other donors;
- Interviews and fact checking with relevant staff from DFID and other donors;
- Quantitative analysis of data from at least 3 sample DFID programmes; calculation of standard unit costs in different contexts and application of proposed methods for calculation of NPV;
- Based on analysis, development of a tool (e.g. a costing/budgeting template) that can be used by DFID advisers and other donors for budgeting for and valuation of evaluations in the future.

Skills Required

Key professional requirements for the contractor are:

- Knowledge of international development evaluation
- Knowledge of the international evaluation literature
- Significant knowledge of and experience in budgeting for evaluations and relevant methodologies from economics, such as CBA and CEA
- Significant experience with quantitative and qualitative data and analysis;
- Significant experience working with evaluation approaches and methods;
- Publication record in relevant topics areas
- Access to a network of evaluation specialists and policy stakeholders;

Timing of the study: Expected start date is 01 November, 2014 - Expected Final Report is 28 February, 2015

The study should take a maximum of 25 days from the start of the contract.
The study team is expected to manage their inputs but an indicative project timetable is given below:

- Up to 4 days FTE for inception report;
- Up to 15 days FTE of desk-based research, interviews and meetings with relevant individuals, analysis of findings, checking results and formulating final recommendations;
- 5 days FTE of writing the report (including interim discussions with DFID);
- 1 day FTE of dissemination and communication activities.

Contractual issues

The project is contracted by DFID and is accountable to DFID. The report should credit DFID for its contribution to the project. DFID will provide a logo for use in the report.

6.2 Case studies

6.2.1 Case Study I - Zambia Social Cash Transfer evaluation – Case Study and Cost Benefit Analysis

1. Background and policy context - *Cash transfers in Zambia*

Zambia has been operating publically funded social safety net schemes since the 1950s; these predominantly included one-off in-kind transfers. The government's ability to fund these varied with its fiscal position – largely one in decline in an extractives-led economy.

A food crisis in the early 2000s focused the need to expand social protection, and improved economic conditions, including HIPC debt-relief, better placed government to do so. A draft National Social Protection Strategy was absorbed into the fifth National Development Plan (NDP) (2006-2010). Social protection maintained its place in the sixth NDP (2011-2015).

Donor involvement with social protection in Zambia started in 2003, with GTZ funding a pilot in Kalomo, through a government scheme. DFID funded this pilot from 2005. Between 2005 and 2010 further pilots were implemented, targeting criteria such as old age, urban populations and district-level scale-up.

The various models were consolidated into the Social Cash Transfer (SCT) programme in 2010, designed by the Ministry of Community Development and Social Services[88] and Cooperating Partners, DFID and UNICEF. The SCT – designed to reduce extreme poverty and intergenerational transmission of poverty - had three targeting models: the Child Grant (for mothers with children under 5 years old) in poorest districts, the Multiple Categorical Grant (targeting female headed households with orphans or elderly members in less poor districts and the 'inclusive model' targeted towards the poorest 10% that were labour constrained in other districts. The SCT was designed to scale-up, aiming to reach 100,000 households across 15 districts in 2015. Funding was joint between government and donors, whose contributions would decrease over the period.

2. Evaluation design and findings

The evaluation was conducted by the American Institutes for Research (AIR) for the government of the Republic of Zambia, under contract to UNICEF, with funding from the Cooperating Partners including DFID.[89]. There were three main reasons for why the evaluation was commissioned in the first place. First, risk attached to the SCT played a role. Cash transfer programmes can be inefficient or ineffective due to poor targeting, high administration costs and losses in the transfer process. Where there is risk of failure or under-performance evaluation provides the basis for modifying or stopping programme extension and thus avoiding costs that would otherwise have been incurred. Second, and related to this point, several senior Zambian ministry officials made an evaluation a condition for their financial and political support for the programme. Third, there was evidence of poor performance in previous cash transfer schemes in Zambia and variable results especially from targeted and conditional cash transfer schemes elsewhere[90]. This indicated that reviews could improve design, delivery and cost-effectiveness. With cash transfers, important decisions need to be made on the target households, the size of the monthly transfer and any conditionality.

[88] Now: Zambian Ministry of Community Development, Mother and Child Health (MCDMCH)?
[89] This is a direct quote from the AIR 24 month report.
[90] Slater R (2011). Cash transfers, social protection and poverty reduction. *Int J. Soc Welfare*, 20 250-259.
Farrington J and Slater R (2006). Introduction: cash transfers: panacea for poverty reduction or money down the drain? *Dev. Policy Review* 24 499-511

The evaluation is of the Child Grant social cash transfer programme (CGP), which targets vulnerable households in three most disadvantaged Districts, which have child under the age of five. They receive ZMW 60 / month, regardless of household size. This amount is deemed by the MCDMCH to be sufficient to purchase one meal a day for everyone in the household for one month. The evaluation is a randomised controlled trial design (RCT) involving 2,515 households (14,565 people), randomly assigned to treatment or control populations. It commenced in 2010, and randomisation was conducted in close cooperation with the Ministry. The evaluation estimates programme impacts on individuals and households using a differences-in-differences model that compares change in outcomes between baseline and follow-up and between treatment and control groups.

The three-year evaluation assesses impacts on expenditures, poverty, food security, children under age 5, children older than 5, and the economy. There were one impact indicator (extreme poverty gap) and six outcome indicators (food security, education, health, child wellbeing, women's empowerment, improved livelihoods). The evaluation also examined the programme implementation, and concluded that the Ministry had done so successfully, disbursing the cash transfers to the right beneficiaries in a timely, cost effective manner. The main findings from the 24 month follow-on survey round (which was the critical report for the scale up decision) were:

Summary of Impacts of 24-month SCT Evaluation Report (results that are statistically significant at .05 level)	
Supplement and not replace household income	Increase of ZMW 15 in monthly per capita consumption expenditure Reduction of 11 % in poverty gap and squared poverty gap
Increase in the number of households having a second meal per day	Increase of 8% in HHs with 2+ meals per day Increase of 22% in proportion of children aged 6-24 months receiving minimum feeding requirements
Reduce the rate of mortality and morbidity of children under 5	Reduction in diarrhoea of 5%
Reduce stunting and wasting among children under 5	Increase in weight for height of 0.196 z-scores among children aged 3-5 years.
Increase the number of children enrolled in and attending primary school	- (no statistically significant effects)
Increase the number of HHs owning assets such as livestock	Increase of 21% in HHs owning any livestock

3. **Uptake pathways: scale up based on the SCT evaluation results and other factors**

Shortly after the publication of the mid-term evaluation results (see above) in 2013, the Zambian Government decided to scale-up the SCT. Parliament approved a budget which tripled the 2014 SCT target from 60,000 to 190,000 households. It increased the government's contribution to the SCT budget by a factor of 8, from 17.5 million Zambian Kwacha (ZMW[91]) in 2013 to ZMW 150 million for 2014. The causality underlying this very

[91] Approx. ZMW 10 : £1

large scale-up decision is central to this paper. In addition to the evaluation results itself, a number of political and context factors are aligned behind this decision.

The Patriotic Front (PF) government elected in late 2011, mentioned in its 2011-2016 Manifesto to adopt a social protection policy and increase budgetary allocations to social protection. It furthermore adopted a new policy framework on poverty reduction (National Social Protection Policy). At the same time, the new government also removed fuel and some maize subsidies shortly before the SCT scale up. In its official communication the President's office repeatedly underlined that this removal would be an opportunity to channel resources to pro-poor programmes.

These aspects all created a context in which scaling up SCT became a real possibility. The Zambian Government stated in its 2014 budget that the scale up of the programme was a shift from 'poorly targeted subsidy programmes where beneficiaries were not the intended poor' to 'better designed social protection programmes such as the social cash transfer scheme that has been successfully piloted', referring explicitly to the evaluation. Key actors believe that the release of rigorous and credible evidence from 24 month round of the CG-IE came at the right time to finally convince the Cabinet Office and the Ministry of Finance to scale-up the budget to the SCT programme. Although the timing of the political factors was not the basis for timing the release of the evaluation findings, once the alignment of factors became apparent, a policy brief was produced to capitalise on the moment. At the same time, a World Bank review of Cash Transfers was also released and contributed to the uptake. The Vice-President has described the scale up as an example of evidence-based policy making. Each of the following evaluation-related factors were important in driving the scale up decision:

- **Quality and rigour:** The SCT itself attained a greater standing as an intervention because it was being rigorously evaluated – it became a more credible candidate for scale-up. This was of course strengthened by the CG-IE showing significant positive impacts could be attributed to the SCT.
- **Communications;** The RCT design and multi-annual surveys, and clear and targeted communication of findings were well received by different Government audiences. A number of senior decision makers in government were medical doctors by training, for whom the RCT particularly resonated. The focus of the CG-IE on assessing the poverty impact of cash transfers also meant it was relevant to audiences across government, rather than just in the Ministry of Community Development. A policy brief was produced to emphasise that the SCT programme has a demonstrable impact on poverty and can be scaled-up.
- **Timing:** The findings from the CG-IE – specifically the 24 month follow-up round – were released at a strategically important time (see political factors above), which helped justify scaling up the SCT programme to improve the poverty-targeting of government social protection schemes.

4. Costs and benefits

The following sections provide detailed analysis of the costs and benefits of the SCT evaluation, focussing on the costs to DFID (evaluation) and GoZ (programme), and benefits to Zambians below the poverty line.[92] The analysis is based on a simplified cost-benefit analysis (CBA) of the social protection programme itself, based on distributional weights.

[92] Costs: The evaluation was original planned to run for three years. It was extended to five years for logistical and other reasons. The five year budget was just under £5.46m. The baseline 2011 and 2014 rounds each cost over $2m; rounds in 2012, 2013, and 2015 cost $1.12m and $1.68m.
This analysis relates only to the evaluation costs, but it worth noting that the SCT programme cost over 10 years (2010-2020) is £55.3m, with DFID investment of £38m (69%).

This CBA does not attempt to assess financial value, rather it tries to estimate the subjective welfare increase of the beneficiary households, net of any programme costs. Distributional weights are based on the intuition that an additional pound of income is worth more to someone on a low income than on a high income.

Measuring SCT programme benefits: The total value of cash transfer programme is multiplied by the distributional weight to calculate the total programme benefit. The total programme cost is calculated by the cost of the cash transfers plus any costs incurred in administering the payments (in this case including a large component for setting up the systems, as these were being introduced for the first time). What value should the distributional weight take? In principle this can be calculated by comparing the income of recipients with the income of taxpayers – the larger the difference, a greater distributional weight can be justified. For this analysis, we have used several reference rates –taken from the cost-benefit analysis of the social protection programme carried out in 2013 and the Green Book. We also calculate the break-even distributional weight.

Assumptions	**Value**	**Description**
Distributional weight 1	2.69	CBA version 1
Distributional weight 2	4.76	CBA version 2
Distributional weight 3	1.52	Break-even value
Distributional weight 4	2.00	Green Book - bottom quintile
Discount rate	10%	DFID standard (alternative = 12%)

Calculating the evaluation NPV: We calculate the net benefit of the evaluation by comparing the benefit of the scaled-up SCT to a non-scale up scenario. We then factor in the cost of conducting the evaluation of the programme to calculate a "*return to evaluation*", or a net present value (NPV) of the evaluation. In addition to the assumptions around the measurement of the benefits (distributional weights), discount rate (10%) and timeframe (costs and benefits between 2013 – 2020), we also need to make an assumption about a counterfactual scenario. That is, a) how much would have been spent on the SCT without the scale up decision ("business as usual"), and b) what would have happened with the *difference* between the government funds that are now being spent on the scaled up SCT programme from 2013 – 2020 and the SCT funds for the business-as-usual scenario 2013 - 2020. We can make relatively credible assumptions for a), based on an earlier government budget predictions. For b), we have created two scenarios, one of which assumes that in the absence of the evaluation, the government would have continued the SCT programme without scaling up, and would have used remaining funds on programmes in which 80% of the spending would have reached the poorest. In this case, the return to the evaluation turns out to be between $15m – $36m, depending on the distributional weight used. In the second, more generous, scenario, we assume that GoZ would have spent the remaining funds on even more effective pro-poor programmes that would themselves have had a multiplier effect (distributional weight) of 1.1. In this case, the NPV of the evaluation ranges between $100,000 and $15m.

The above analysis does not assess whether the cost of administration, or the cost of the evaluation, is reasonable. Nor does it ask whether the evaluation was necessary in order to trigger the scale-up of the programme – this is an assumption of the analysis, based on the qualitative narrative above. It may be that the evaluation was more costly than necessary. Equally, it may be that the evaluation contained information that was useful in other ways, that are not considered by the analysis.

5. Conclusion and learning about ex-post valuation of evaluations

In conclusion, it should have become clear that an ex-post CBA can be helpful to calculate a *range* of NPVs that can give a rough estimate of the proportionality of an evaluation, and can help to demonstrate the impact an evaluation has had. However, there are three key challenges that emerge from the above analysis.

First, attributing changes in policy or programme design to an individual evaluation is only possible through qualitative narrative analysis, as outlined above. As in the SCT example, there will always be a host of factors contributing to social impact, and the quantitative analysis cannot account for these fine grained nuances. Second, measuring the benefits stemming from an evaluation will be difficult in many sectors, such as in the SCT example above. Beyond issues around quantification and how far down the results chain one should look, another question is about benefits and costs *to whom* we should take into account. Third, calculating the NPV of an evaluation, one has to make assumptions around *a counterfactual scenario*, which in many cases will be arbitrary. Only in few occasions will it be clear *what would have happened in absence of the evaluation*, both in terms of costs and benefits.

However, as demonstrated in the second case study below, the extent to which these three challenges are actually relevant when calculating net benefits of evaluations differs, depending on the type of evaluation. If evaluations analyse similarly effective interventions with differences only in cost-efficiency we can overcome most of the measurement and counterfactual challenges and more conveniently calculate cost savings, as outlined below.

6.2.2 Case Study II - Achieving Net Savings Through Evaluations of Development Interventions - Evidence from an Insecticide-treated bed nets Evaluation in Zambia

1. Background and policy context - Malaria in Zambia

Malaria is a major public health challenge in Zambia, accounting for 36% of hospitalisations and outpatient department visits, 8-14% of low birth weight babies, 3-8% of all infant deaths, and up to 20% of maternal mortality.[93] Insecticide-treated bed nets (ITNs), are a cost-effective way of decreasing malaria incidence. Consistently sleeping under an ITN has been shown to decrease all-cause child mortality by 17-29%, and larger community-wide gains can be realised if a critical number of households utilise ITNs.[94] In 2014, the Zambian government was planning to distribute 6-7 million ITNs, but only limited information was available on the most efficient distribution and hang-up approach to optimise ITN ownership and use in Zambia. A door-to-door distribution strategy had been previously used throughout Zambia, but was highly challenging and costly in terms of time, supervision costs, and volunteer work burden. In order to inform decisions about using alternative methods, Zambia's Ministry of Health (MOH) and National Malaria Control Centre (NMCC) requested an evaluation to assess the cost-effectiveness of community fixed-point distribution of ITNs compared to the current door-to-door strategy. An evaluation was conducted in collaboration with the Demand-Driven Evaluations for Decision (3DE)[95] program and World Vision Zambia.

The 3DE pilot

BACKGROUND: Despite the growing recognition of the importance of an evidence-based approach to global health policy-making, the gap between evaluation evidence and policy persists, particularly in low-resource settings. The Clinton Health Access Initiative's (CHAI) Demand Driven Evaluations for Decisions (3DE) pilot, launched in Uganda and Zambia in 2012 with funding from the UK Department for International Development (DFID), is pioneering a new evaluation model to support ministries in the health sector with evidence-based decision-making by using rigorous impact evaluations in a demand-driven, rapid and efficient way.

AIM: The primary aim of the 3DE project is to work alongside health policy makers to generate reliable impact evidence that meets the ministries' needs and is directly used to catalyse implementation of cost-effective health interventions. The 3DE approach combines a number of important aspects for successful evaluations, namely the integration of evaluation into the programme and planning cycle, timely delivery and the involvement of stakeholders. Since 2012, the 3DE programme has completed five demand-driven impact evaluations on diverse health-related topics, including the above-mentioned ITN-evaluation.

2. Evaluation design and findings

The 3DE evaluation tested an approach that involved distributing ITNs from a pre-defined fixed point in a community such as a clinic, school or church, with community health workers (CHWs) visiting households after a given time period to hang unused ITNs. In the evaluation, this community fixed-point distribution strategy was tested in three locations in Rufunsa District, using randomised and observational evaluation techniques to measure important indicators such as household attendance, ITN use, ITN retention, and CHW time required.

[93] National Malaria Control Centre (2015). Malaria Control in Zambia. nmcc.org.zm/malaria_control.htm
[94] President's Malaria Initiative. Fact Sheet on Insecticide-treated Mosquito Nets (ITNs). <http://www.fightingmalaria.gov/news/pressreleases/itn_facts.html>
[95] See http://devtracker.dfid.gov.uk/projects/GB-1-203504/ for more information and background documentation about the 3DE pilot.

The evaluation utilised two interventions: The first intervention was community fixed-point distribution of ITNs. All communities in the evaluation received this type of distribution and pre-registration data from a household survey conducted by the Rufunsa District Health Office (DHO) and World Vision were used to determine the households invited to each distribution and number of ITNs allocated for each household. Observational methods were used to measure feasibility of the intervention, particularly through household attendance rates. The second intervention involved door-to-door visits by CHWs to hang up unused ITNs. Because CHWs were not required to carry ITNs during these visits or to hang all of the ITNs, the visits were expected to be less time-consuming than the traditional door-to-door distribution. CHWs were trained on malaria messages, ITN hang-up techniques, and basic data collection methods. The timing of CHWs visits ranged from 1 to 17 days after distribution or no CHW visit at all, and the timing of visits was randomly assigned. CHWs collected data about self-installation at the time of their visit. A household survey was conducted at 7-11 weeks following the point distributions to measure ITN use and retention rates. 577 households from 3 communities were included in the analysis.

This evaluation indicated that the community fixed-point distributions achieved ITN use and retention rates comparable to a well-run ITN door-to-door distribution while achieving significant time and cost savings.[96] 95% of households in pre-registration survey attended the fixed-point distribution events or sent a representative. At 7-11 weeks after the distributions, 91% of distributed ITNs were found in households and 77% of all observed sleeping spaces were covered by an ITN. When volunteers visited households following the fixed-point distributions, many families had already hung their ITNs, saving time and effort on the part of the volunteers. Specifically, 46% of distributed nets were self-installed 5-7 days after distribution and 73% of distributed nets were self-installed 10-12 days after distribution. According to a model based on operations in Rufunsa, community fixed-point distribution could reduce the CHW time required by approximately 35%. Based on this evaluation and data available from other door-to-door distribution campaigns, it was estimated that the total cost for the 2014 Mass Distribution Campaign, including CHWs and supervision costs, would be ZMW 17,500,000 (approx. USD $3 million) for door-to-door distribution and ZMW 11,500,000 (approx. USD $2 million) for fixed-point distribution.

3. **Uptake pathways: scale up based on the ITN evaluation results**

Shortly after the findings of the ITN evaluation were disseminated, the MOH decided to scale-up the community fixed-point approach for the 2014 ITN distribution round. Several key factors contributed to this decision. As part of the 3DE pilot (see box above), CHAI had closely engaged with key officials throughout the ITN evaluation process. Firstly and most critically, the initial evaluation question was sourced from and shaped together with the MOH ensuring ownership and increasing the likelihood of uptake of results from the evaluation. Secondly, key staff from the MOH as well as key implementing partners were represented on the evaluation team as co-investigators. This ensured MOH and stakeholder input into the evaluation design at protocol stage as well as its implementation in the field. Thirdly, the evaluation design was presented in the key ITN policy making space - the ITN Technical Working Group (TWG) - for stakeholder review and feedback. This process ensured that once results were presented in the same TWG a few months later, they were fully accepted by stakeholders and could thus be used as a basis to inform a decision regarding ITN distribution strategy. Additionally, once the decision to scale-up the approach was made at national level, the strategy was presented at several planning meetings ahead of the distribution campaign with subnational level representation. This helped to ensure that the implementing entities at subnational level were sensitized concerning of the policy change and helped to ensure its implementation.

[96] STEPS OVC (2013). Report. House to house distribution of insecticide-treated nets in nine Zambian districts, 2011-2012

4. Costs, benefits and conclusions

The following sections provide detailed analysis of the costs and benefits of the ITN evaluation.

Cost of Evalaution		
Item	*Cost (USD)*	*Source*
Evaluation overhead and admin	**$ 74,004**	3DE evaluation records
Salaries for data collectors	**$ 25,622**	3DE evaluation records
Transportation to sites	**$ 29,732**	3DE evaluation records
Misc. costs	**$ 6,620**	3DE evaluation records
Total cost of evaluation	**$ 135,978**	Calculation

Table A1 – Cost of Evaluation

Based on the evaluation results, costs of the door-to-door distribution were estimated at $1.13 per ITN, whereas the fixed-point distribution with delayed hang-up resulting in comparable hang-up rates costs $0.82 on average (see Table A2).

Cost of ITN distribution methods		
Item	*Cost (USD)*	*Source*
Cost per ITN by door-to-door distribution	$ 1.13	Costing exercise based on 3DE evaluation results
Cost per ITN by fixed-point distribution w/ delayed hang-up	$ 0.82	Costing exercise based on 3DE evaluation results
Savings per ITN of using fixed-point distribution with delayed	**$ 0.30**	Calculation

Table A2– Cost of ITN distribution methods

As outlined above, following the dissemination of the evaluation results the NMCC released a policy memo providing guidance that districts could choose whether to use door-to-door or community fixed-point distribution in the 2014 ITN distribution, depending on local characteristics such as population density and CHW availability. In total, just under 5,000,000 ITNs were distributed in Zambia following the policy memo, and nearly 20% of these via the fixed-point distribution approach. This resulted in absolute savings of more than $291,000 compared to a door-to-door-only distribution. The primary source of savings in the community point distribution model was the time spent by community health workers fetching and hanging ITNs in households. Subtracting the initial evaluation costs, **the evaluation led to net savings of just under $155,269 in 2013/14** (see Table A3). If planned future distribution rounds would be taken into account, the net savings would be accordingly higher.

Savings / Benefits of Fixed-Point Distribution Scale-up		
Item	***Value***	***Source***
Total number of ITNs distributed in 2014 distribution	**4,939,989**	Calculation
Door-to-door distribution	3,982,479	Primary data collection from districts
Fixed-point distribution w/ delayed hang-up	957,510	Primary data collection from districts
Cost for actual distribution method mix used	$ 5,268,190	Calculation
Cost for hypothetical CFP distribution level (50% with delayed hang-up)	$ 4,808,135	Calculation
Cost if done by door-to-door distribution only	$ 5,559,438	Calculation
Savings due to incorporation of fixed-point distribution	**$ 291,248**	Calculation
Net savings (minus evaluation costs)	**$ 155,269**	Calculation

Table A3 – Savings and Benefits in the 2014 ITN distribution

Analysing the health benefits of ITNs, we find that up to 81,500 under-5 deaths were prevented by the 2014 ITN distribution, which corresponds to 2,689,000 DALYs saved (see Table A4).

Health Benefits of ITNs		
Item	***Value***	***Source***
% of ITNs retained at 6 months	86%	3DE evaluation results, based on household survey self-report (79% based on visual inspection) - no difference for CFP and D2D
% of sleeping spaces covered at 6 months	80%	3DE evaluation results, no difference for CFP and D2D
% of ITNs hung in household at 6 months	64%	3DE evaluation results, no difference for CFP and D2D
Number of all-cause under 5 deaths prevented per year per 1000 children	5.5	Lengeler C, 2004
Average lifespan of ITN (years)	3	Pulkki-Brännström et al, 2012 WHO, 2010.
Number of all-cause under 5 deaths prevented by 2014 ITN distribution	81,510	Calculation
Number of DALYs saved by averting 1 under-five death	33	Pulkki-Brännström et al, 2012 WHO, 2010.
Number of DALYs saved by 2014 ITN distribution	2,689,824	Calculation

Table A4 – Health benefits of ITNs[97]

[97] References in the tables: Lengeler C. (2004). Insecticide-treated bed nets and curtains for preventing malaria. In Cochrane Database of Systematic Reviews. Chichester, UK: John Wiley & Sons, Lt
http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000363.pub2/abstract

As highlighted in Table A5, this translates into costs per DALY saved of $9.36 for the 2014 distribution round, with net savings of $0.05 per DALY saved using the fixed-point distribution method.

Cost-effectiveness of Evaluation		
Item	***Cost (USD)***	***Source***
Cost per ITN (excluding distribution)	$ 4.00	Pulkki-Brännström et al, 2012
Cost per DALY saved by ITNs distributed using D2D distribution	$ 9.41	
Cost per DALY saved by ITNs distributed using actual mix of D2D/CFP distribution, plus the cost of the 3DE evaluation (eval costs only)	$ 9.36	
Hypothetical for future) Cost per DALY saved by ITNs distributed using 50/50 mix of D2D/CFP distribution, plus the cost of the 3DE evaluation (eval costs only)	$ 9.18	

Table A5 - Cost-effectiveness of evaluation

Pulkki-Brännström AM, Wolff C, Brännström N, Skordis-Worrall J. (2012. Cost and cost effectiveness of long-lasting insecticide-treated bed nets - a model-based analysis. www.resource-allocation.com/content/10/1/5
WHO (2010). Chapter 4 Vector control. In World Malaria Report 2010. Geneva: WHO. http://www.who.int/malaria/world_malaria_report_2010/en/

6.3 Budgeting evaluations

DFID-internal annex. Not for external publication

6.4 Guidance on data collection for use in ex-post CBA of evaluations

This is a basic guide to what data would need to be collected to enable an ex-post CBA or CEA calculation of an evaluation. In general, the CBA/CEA requires a listing of costs and benefits by year over the life of the 'project'[98] (normally **the time period over which benefits occur**). These should be incremental costs and benefits as compared with a counterfactual (no evaluation). Costs can normally be recorded in monetary terms whereas benefits are best recorded as quantities, leaving valuation of the quantities until the analysis (CBA) stage. Overall, it is suggested that DFID introduces the idea of developing an ex-post CBA/CEA in a number of its evaluations, particularly larger, multi-annual impact evaluations. Contractors would be tasked with working with EVD during inception to develop CBA templates for their evaluation, progressively collecting the necessary data to be able to conduct the ex-post CBA after completion.

Costs

This should include all costs associated with the evaluation. It may be appropriate to include an element of 'overhead' costs but this depends on whether or not the evaluation is treated as incremental (with a set of overheads that would be incurred anyway) or one incurring additional or allocated overheads.

To be included in the measurement of costs:

- Evaluation costs tabulated by year and categorised by donor, currency, type (usual disaggregated accounting categories mainly to ensure that all costs are included).
- Evaluation costs should be specified by type: consultancy fees, associated staff and other costs to undertake the evaluation.
- Transaction costs in implementing the evaluation unless included elsewhere.

As noted in EVD's 2013-14 Annual Report: "*One of the challenges in tracking evaluation costs is that the evaluation budgets estimated during a planning process may differ from the actual contract costs for the evaluations once they are commissioned*". Thus it is important that **actual costs** are accurately recorded. DFID will need to agree a standard method for recording or estimating staff FTEs dedicated to individual evaluations, and similarly for associated expenses. Evaluation contractor costs should be simpler to obtain, through ARIES records of their invoices against the evaluation project code, but can be complicated in the case of multi-donor evaluations.

Benefits

A qualitative specification of the benefits expected from the use of the evaluation's evidence and their time scale. One aim of the evaluation will be to quantify these benefits. As part of their design, evaluations should be required to specify the benefit measures that will be used for CBA (or CEA) calculations. These may be outcomes or impacts. Likewise, in their designs, evaluations should be required to detail how the specified benefits will be valued.

[98] In the particular case of ex-post CBA of *evaluations*, the 'project' is made up of the evaluation for the cost side (the cost occur within the life of the evaluation) and a number of users of the evaluation who realise its benefits. These may be the evaluand itself, benefitting internally from the evaluation evidence to improve its performance, or other programmes (and policies) who take up and scale the evidence, often after the evaluation concludes.

In broad terms, for impact, the aim is to measure the change in welfare of the target population using direct measures or indicators. Since benefits may be in terms of improved employment, incomes, health, education, environment, security etc, it is impossible to specify generally what should be measured since each category of investment and programme will have different requirements.

It is suggested that the best approach is to take example interventions and list what measurements might be appropriate as a basis for CBA/CEA – to develop a menu of the most likely impact (and outcome) measures that could be monetised for CBA of evaluation or quantified for CEA (e.g. DALYs in health). Benefits anticipated over a long time horizon (e.g. in educational attainment) and those that are not valued in markets (e.g. environmental improvements) are especially difficult to measure since evaluations may be too short-term to include these in any precise way or require specialised measurement techniques.

Contribution/attribution: As well as collecting data on longer terms benefits, it is also necessary to track whether and how the evaluation is being used by policy makers. In order to undertake the CBA a case will need to be put forward that shows that the evaluation did in fact contribute to a decision which in turn led to specific development outcomes being achieved. Without a compelling contribution story for the evaluation the CBA analysis will have little credibility. As shown by the case studies in Annex 6.2, attributing changes in policy or programme design to an individual evaluation is only possible through qualitative narrative analysis. As in the SCT example, there will always be a host of factors contributing to social impact, and the quantitative analysis cannot account for these fine-grained nuances.

Counterfactual: Furthermore, as outlined in the case study, when conducting a CBA/CEA of an evaluation, one has to make assumptions around a counterfactual scenario, which in many cases will be difficult to justify. Only in few occasions will it be clear what would have happened in absence of the evaluation, both in terms of costs and benefits. It is important to consider which data are needed to construct a counterfactual scenario for the CBA/CEA of evaluations. However, as demonstrated in the case studies, the extent to which the above challenges are actually relevant when collecting data for CBA/CEA of evaluations differs, depending on the type of evaluation. If evaluations review similarly effective interventions with differences only in cost-efficiency we can overcome most of the measurement and counterfactual challenges and more conveniently calculate cost savings, as outlined in the Zambia ITN case.

Benchmarks

Commissioning organisations often use a benchmark to specify considerations in making a budget estimate: "*common budget estimates range between 5 – 20% of program costs*"[99]. This remains a broad range, and there are a number of limitations in using this approach. It is often not clear whether the '5-20%' range relates only to external evaluation, or if it includes internal monitoring activities and commissioning costs. Further, this range can lead to some unrealistic cost estimates: 20% of a £100m programme would be an excessive amount for an evaluation, while 5% of a

[99] BetterEvaluation: http://betterevaluation.org/plan/manage_evaluation/determine_resources

£100,000 programme would not be sufficient to conduct an evaluation. For this reason, these figures are not recommendations. The figures need tailoring according to the characateristics of the evaluand and the proposed evaluation.

Although actual evaluation budgets are, mainly for commercial sensitivity reasons, not widely available[100], there is a range of sources providing benchmark ranges. For example, the EC INTERACT service suggests in its evaluation handbook that a range of 1% to 10% of programme resources may be allocated to evaluation (Box 4)[101].

> "*for large-scale relatively routine programmes the budgets required for evaluation will be a small proportion of the programme resources (normally less than 1%). On the other hand, for interventions that are relatively innovative and pilot in character and where evaluation has a strong learning and participatory aspect the costs are likely to constitute a relatively high proportion of programme resources (up to 10%). There are instances where up to 5% of programme budgets have been devoted to evaluations that are effectively part of the management's implementation strategy, for example, where evaluation includes a strong formative element intended to assist managers and stakeholders with their work. It is useful to indicate a minimum and maximum budget*."

Box 4. Proportion of Programme Budget Used for Allocation

[100] Even evaluation databases, such as 3ie's Registry for Development Impact Evaluations (RIDIE), though improving the transparency of design and analysis in impact evaluations, are silent on the matter of cost.
[101] EC INTERACT: http://wiki.interact-eu.net/pages/viewpage.action?pageId=23756932

6.5 Examples of schematic presentations to summarise case studies

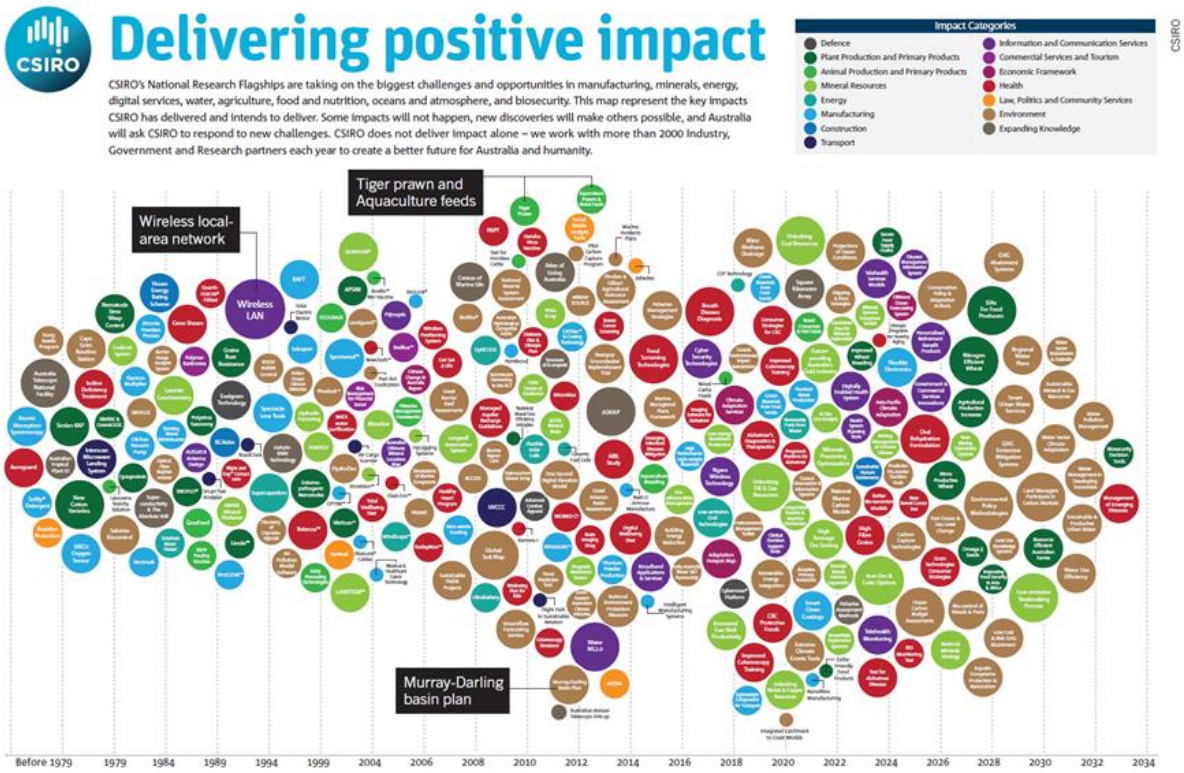


Figure 7. Map of CSIRO impact case studies

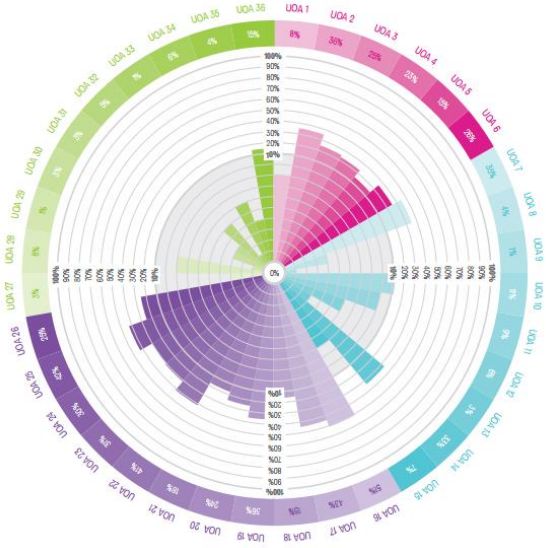


Figure 8. REF Impact Wheel - 'Informing government policy'[1]

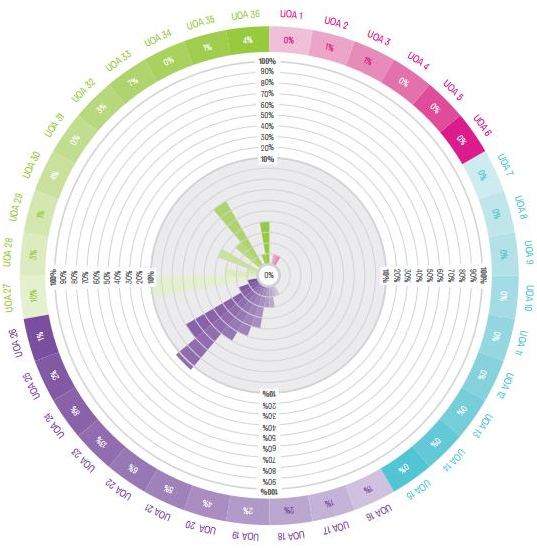


Figure 9. REF Impact Wheel - 'Women, gender, and minorities'