Overseas Development Institute

Improving impact evaluation production and use

Nicola Jones, Harry Jones,
Liesbet Steer and Ajoy Datta

Working Paper 300

Results of ODI research presented in preliminary form for discussion and critical comment

Working Paper 300

Improving impact evaluation production and use

Nicola Jones, Harry Jones,
Liesbet Steer and Ajoy Datta[1]

March 2009

Overseas Development Institute
111 Westminster Bridge Road
London SE7 1JD

[1] Valuable research assistance from Cora Walsh, Carlotta Tincati, Anna Caffell, Laura Gisby and Hannah Marsden is gratefully acknowledged.

ISBN 978 0 85003 899 6
Working Paper (Print) ISSN 1759 2909
ODI Working Papers (Online) ISSN 1759 2917

Contents

Tables, boxes and figures

Acronyms

3IE	International Initiative for Impact Evaluation
ADB	Asian Development Bank
AIDS	Acquired Immunodeficiency Syndrome
ALNAP	Active Learning Network for Accountability and Performance in Humanitarian Assistance
ARV	Antiretrovirals
BRAC	Bangladesh Rural Advancement Committee
CCT	Conditional Cash Transfer
CEF	Comprehensive Evaluation Framework (EC)
CGD	Center for Global Development
CGIAR	Consultative Group on International Agricultural Research
CI	Catalytic Initiative to Save One Million Lives
CMI	Chr. Michelsen Institute
CONEVAL	National Council for Evaluation of Social Development Policy (Mexico)
DAC	Development Assistance Committee
DFID	UK Department for International Development
DIME	Development Impact Evaluation Initiative
EC	European Commission
EES	European Evaluation Society
EF	Evaluation Framework
EU	European Union
GAVI	Global Alliance for Vaccines and Immunisation
GPOBA	Global Partnership for Output-based Aid (World Bank)
HIV	Human Immunodeficiency Virus
IADB	Inter-American Development Bank
ICT	Information and Communication Technology
IE	Impact Evaluation
IEG	Independent Evaluation Group (World Bank)
IFPRI	International Food Policy Research Institute
IHP	International Health Partnership
ISS	Immunisation Services Support (GAVI)
JBIC	Japanese Bank for International Cooperation
JICA	Japan International Cooperation Agency
J-PAL	Abdul Latif Jameel Poverty Action Lab
MDG	Millennium Development Goal
NGO	Non-governmental Organisation
NONIE	Network of Networks on Impact Evaluation
Norad	Norwegian Agency for Development Cooperation
NRM	Natural Resources Management
ODI	Overseas Development Institute
OVE	Office of Evaluation and Oversight (IADB)
PART	Programme Assessment and Rating Tool
PEPFAR	President's Emergency Plan for AIDS Relief
PRA	Participatory Research Approach
PREM	Poverty Reduction and Economic Management
PROLINNOVA	Promoting Local Innovation in Ecologically Oriented Agriculture and NRM
PRSP	Poverty Reduction Strategy Paper
RAPID	Research and Policy in Development (ODI)
RBF	Results-based Financing
RBM	Results-based Management
RCT	Randomised Control Trial
SIEF	Spanish Trust Fund for Impact Evaluation
SWAp	Sector-wide Approach
UNDP	UN Development Program
USAID	US Agency for International Development

Executive summary

Introduction

The past five years have seen a proliferation of impact evaluations (IEs) by development agencies across the globe. This report was commissioned by the UK Department for International Development's (DFID's) Evaluation Department to inform discussions on impact evaluation production and use within the Network of Networks Impact Evaluation Initiative (NONIE). It builds on an initial scoping study prepared for DFID which made recommendations on improving IE production and use, focusing on clustering, coordination, knowledge management, capacity strengthening and communication and uptake. This the report goes further by expanding both the literature review and the annotated database of IEs, as well as honing in on specific dynamics of IE production across sectors.

Methodological approach

An initial analysis of existing IE literature was combined with an annotated database of IEs undertaken in developing countries. Based on this review, a set of hypotheses was established, which were tested in a variety of sectoral case studies. These studies were illustrated by highlighting productions paths and obtaining input from key informants. The findings were compared and contrasted with present conclusions and implications for policy audiences.

Key findings from sectoral analyses

A focus on sector-specific histories and dynamics of impact IE production, communication and use dynamics revealed a number of important similarities and differences. Similarities included a growing recognition of the need to approach IEs as part of a broader monitoring and evaluation system; the importance of involving multiple stakeholders in the evaluation process to promote uptake; and the utility of exploring alternative methods to assess impact.

Key differences appeared to be starker and were found in a number of areas. First, a longer history of IEs in health and agriculture/natural resource management (NRM) sectors has meant these sectors have a broader knowledge base from which to draw, although they diverge in the extent to which this knowledge is actually used.

Second, there is a greater recognition in health and to a lesser extent agriculture/renewable and natural resources sectors that IEs are strongly suited to providing robust evidence on a range of key questions in the field. This recognition is growing in the social development sector, but views in humanitarian, infrastructure and results-based aid sectors are more cautious about the relevance of certain methods.

Third, there is strong interest and practice of methodological innovation for IEs in the health and social development sectors, increasingly so in agriculture/ renewable natural resources, but less so in humanitarian and infrastructure sectors.

Fourth, in health, social development and results-based aid, actors recognise that impact often takes several years, with impact in agriculture/NRM and infrastructure taking considerably longer.

Fifth, commissioning of IEs tends to be supply driven in the case of all but the health and social development actors, where there is growing demand from developing country governments, especially in Latin America.

Sixth, communication of findings at the national level appears relatively robust in the health sector, whereas in other sectors this seems to be more limited, and in the results-based aid sector it is still too early to assess.

Seventh, IEs are routinely used in the medical field, and there is growing uptake in the public health field. In agriculture/NRM, however, there is concern that use of findings is hindered by limited attention in IEs to broader context and programmatic variables. In the humanitarian sector, there is no concern that evaluation results are unduly biasing donor policies, whereas in the infrastructure and rural/urban development sector, there is a general perception that the use of results stops at reporting donors and official accountability objectives.

Finally, in the case of results-based aid, there is recognition that IEs will remain an important tool, but it will be key to assess the spillover effects of this new aid modality on the monitoring and evaluation cultures of other development sectors.

Conclusions and policy implications

Strategic coordination: There is a need for a broader strategic framework for impact evaluation production and use that ventures beyond the level of project interventions and addresses wider policy-level questions and challenges. Clustering initiatives that are well informed, owned and linked with broader evaluation systems and that have the potential of going to scale with technical and political support would make a useful contribution to strategic coordination. A further concern for coordination is replicability. As such, certain regional contexts (South Asia, sub-Saharan Africa) will require support, where investment in development initiatives has been greater than corresponding resources (especially funding and capacity development) invested in IEs.

In terms of methods, a focus beyond narrow debates on specific experimental methodologies and compartmentalised professions toward mixed and plural analytic approaches is necessary. Such nuanced evaluation approaches use contrasting perspectives to better capture the reality of a development context. Appropriate methods could be promoted through increased knowledge sharing and introducing more rigorous quality standard guidelines, while maintaining flexibility.

Funding: A critical variable in shaping evaluation practice is funding policy, and the incentive structures that this creates. Greater attention is needed to promote process changes such as wider consultations, better sequenced and integrated lesson learning processes and closer engagement between implementation and evaluation staff. Funding to encourage the involvement of evaluation staff in disseminating beyond academic realms (although clearly important) to policy and practitioner audiences as well as the media is also identified as a critical area. In terms of content, incentives to ensure publication of negative as well as positive results would promote learning and accountability. Finally, investment in evaluative capacity building, along with replication evaluations and pioneering pilots in diverse contexts, would promote knowledge sharing and learning.

Knowledge management: Agreements on common database formatting, updating and circulation are required to promote greater transparency and knowledge sharing. However, there should also be an increased awareness, not only on content evaluations, but also on how they are used to influence policy and practice. Drawing from the International Food Policy Research Institute's (IFPRI's) impact assessment discussion paper series, NONIE could drive an initiative that funds and coordinates such documentation and analysis.

Capacity strengthening mechanisms: Developing country evaluation capacities will need direct support to avoid the possibility of IEs being only donor-driven or supply-oriented tools. Possible capacity development approaches include learning by doing; support for a community of practice including developing country actors; training workshops for 'educated consumers' of IE; supporting the development of national centres of excellence in IE that can partner with international agencies; peer

review of proposed IE methodologies; and integrating impact evaluations into broader capacity building initiatives on evaluation methods.

There is also a broader need for the creation of a clear decision framework that outlines whether, when and what to evaluate. NONIE could respond by contributing to a decision framework and facilitating progress by summarising suitability and plausibility issues that move current debates toward a common ground.

Improving IE communication and uptake: Best practice examples demonstrate that a central factor facilitating uptake of IEs is stakeholder involvement. This involvement must be brought in at the early stages of the IE process, include the support of high-profile champions and attract political agents interested in learning or using instruments to demonstrate effectiveness. Questions regarding potential utilisation raised by these parties can then be integrated into the design of the IE. Another key factor that facilitates uptake is the dissemination of IEs containing clear policy implications to a wide audience.

There is also a need to acknowledge the role of evaluations in policy transfer, and understand the purpose of building up lessons from interventions as part of a global public resource. Particular ingredients of success are a presence of a critical mass of evaluations; a combination of process and impact information; a drive for technical rigour; and the inclusion of cost data. Finally, the findings generated must be related to their context to ensure that any messages taken up are used appropriately.

1. Introduction

In response to a growing demand to assess the impact of development assistance policies and programmes, the past five years have seen a proliferation of impact evaluations (IEs) by donors and some national governments. Impact evaluations have been undertaken on a wide range of development interventions, from school textbooks to cash transfers, gender empowerment and corruption reduction programmes to infrastructure initiatives. In several cases, such as evaluations on school worming programmes and Mexico's conditional cash transfer (CCT) programme, the findings have been highly influential. However, in general, there has been little analysis of the dynamics of the use of IE findings in policy development and programming.

This report builds on an initial scoping study, 'Improving Impact Evaluation Coordination and Uptake', which set out to make recommendations to the Department for International Development (DFID) Evaluation Department and the Network of Networks on Impact Evaluation (NONIE) about how to improve the production and use of IEs once they had been conducted, focusing on clustering; coordination; knowledge management; capacity strengthening; and communication and uptake. The current study aimed to meet the following objectives:

- To determine how amenable IEs are to different types of projects, programmes and policies;
- To look at various methods used to conduct IEs;
- To assess the dynamics around commissioning, production and delivery of IEs;
- To analyse how IEs are disseminated and communicated;
- To assess use and influence of IEs; and
- To make recommendations to improve the production and use of IEs.

Given what is known about the differential dynamics of the research–policy interface across policy sectors (Pomares and Jones, forthcoming), we pay particular attention to similarities and differences in the patterns of impact evaluation production and use across sectors.

The methodological approach adopted for this study involved the following components:

1. **An analysis of existing web-based and published literature** to provide an overview of key debates, both historical and contemporary, about a number of elements of IE production, communication and use, comprising: i) the suitability of IEs to different projects, programmes and/or policies; ii) methodological issues around IE production; iii) supply and demand characteristics of IEs; iv) communication and dissemination of IEs; v) use and influence of IEs; and finally vi) steps to improve production and use.
2. **The development of an annotated database of IE studies** (drawing on five of the main evaluation databases (Development Impact Evaluation Initiative – DIME; NONIE; Poverty Reduction and Economic Management – PREM; Poverty Action Lab – J-PAL; and Consultative Group on International Agricultural Research – CGIAR), with information on thematic, sectoral and geographic areas, and on methodologies (quantitative, qualitative and mixed methods), followed by an analysis coverage in terms of themes, sectors, geographic areas and methodologies in order to set the scene for more in-depth case studies.
3. **The development of a number of hypotheses** emerging from the review concerning the production, communication and use of IEs.
4. **The testing of hypotheses through case studies** to distinguish similarities and differences of IE production and use dynamics in different sectors. Based on IE density in different sectors (from the annotated database), consultations with the DFID Evaluation Department and other key stakeholders, six sectors were selected: health, social development, renewable and natural resources, humanitarian, rural/urban development and infrastructure, with an additional case study looking at the production and use of IEs in results-based aid initiatives. Each sectoral case study was developed through between three and eight key informant interviews with both producers and users of IEs in both developed and developing countries. In total, 62 key informant interviews were undertaken. However, owing to a number of constraints, coverage

was somewhat uneven, with a greater number of IE producers (such as donors and researchers) in the developed world (North America and Europe) than users, especially in developing country contexts. Subsequent work could usefully explore the views of the latter set of stakeholders.

5. **The development of a synthesis** drawing on the five case studies.

The report is structured as follows:

Section 2 begins with a discussion of IE definitions, before presenting an overview of some of the historical and theoretical issues around: suitability of IEs to different interventions; different IE methodologies; supply and demand characteristics of IEs; how IEs are disseminated; and use and influence dynamics and efforts to improve the knowledge base on IEs to date. Throughout this section, tentative hypotheses about the production and utilisation of IE evidence are formulated and proposed

Section 3 provides an overview of IE coverage by sector, geographical region, methodological approach and implementing agency, drawing on documentary analysis.

Section 4 outlines IE production and use dynamics in six sectors in turn: health, social development, renewable and natural resources, humanitarian, rural/urban development and infrastructure, and the aid sector more broadly.

Section 5 presents a synthesis of the six sectors drawing out key themes and lessons.

Section 6 summarises the study's key findings, concludes and teases out a number of policy implications regarding production, communication and use of IEs.

2. Key issues in the relevance, production and use of impact evaluations

This section outlines a number of key issues around the relevance, production and use of impact evaluations in international development. It draws on a systematic literature review, covering published articles as well as grey literature, and also on semi-structured interviews from the scoping phase of the study (see Appendix 4 for key informants). We identify a number of hypotheses, which we reflect on in the sectoral case study section and also use to shape our comparison of the dynamics of IE production, communication and use across the sectors. An important caveat must be made: as donors and much of the literature focus on experimental and quasi-experimental methods for carrying out IE, this is the main focus of the discussion. This is not meant as a value judgment about such approaches.

2.1 IE: Concepts, methods, nature of the knowledge produced

There is a large amount of literature focusing on concepts of impact evaluation and the various methodologies, with their attendant strengths and weaknesses (e.g. Baker, 2000; Bamberger, 2006; Riddell 2008; Roche 2000). Much of this work is quite technical, and debates are frequently epistemological and highly polarised in nature, representing opposing paradigms of social science and development.[2] We will not attempt to settle these debates here, but will give an overview of the issues in a manner relevant for policy and practice. This section looks at the meaning of impact evaluation, and sets the methodological debates briefly in historical context. It discusses the relevance of the knowledge produced and the key considerations involved when it is feasible to assess impact.

2.1.1 IE: Meaning and methods in context

At the broadest level, debates about 'impact' involve looking at the effects of development interventions on their surroundings. In contrast with asking whether an intervention is doing the right thing, or doing it in the right way, it is about whether it has the right effects. Debates often focus on looking at a particular 'level' of effects, on wellbeing outcomes of beneficiaries, and often involve an evaluation some time after the end of an intervention. Concern about evaluating impact has been rising up development agendas in reaction to increased aid flows and attention to the effectiveness of aid (Prowse, 2007). This, in turn, has led to a move from monitoring and evaluation based on outputs, the immediate goals such as building schools, training nurses or making credit available, to looking at outcomes and impacts, what happens outside the direct work of the programme and contributing to people's lives (Riddell, 2008). It is often related to the broadest 'goals' of development, such as the contribution it makes to reducing poverty.

There are different perceptions of how impact should be evaluated, but currently one approach has dominated donor discourse on IE. In many aid agencies, and in certain schools of evaluation, IE refers to an evaluation that assesses the effects of an intervention using a 'counterfactual', which tends to be assessed with experimental or quasi-experimental methods (for example, ADB, 2006). A counterfactual involves comparing what happened with what would have happened had the project not taken place, or what otherwise would have been true. Experimental designs evaluate the counterfactual by randomly assigning the intervention within a well-defined group and comparing the beneficiaries targeted by the intervention with those who did not (the 'control' group). This way, the differences in outcomes between the groups can be attributed solely to the intervention. Quasi-experimental methods are applicable where the programme was not randomly assigned, using statistical methods (such as propensity score matching) to simulate a control group.

[2] For example, CGD (2006) implicitly assume that experimental IEs are the only types of study that can give reliable information about effectiveness and that anyone would use the method given sufficient resources to carry them out; Smutylo (2001) argues that attribution is very rarely possible using any method, and that it is counterproductive to look at impact for learning and accountability purposes.

This view of quantitative, experimental IE being viewed as a 'gold standard' is quite widespread (EES, 2007). For example, the World Bank definition specifies that a counterfactual must be used, and pays little attention to the possibility of assessing it using qualitative methods (IEG, 2006). There is a high degree of scepticism among proponents of this approach as to the extent to which qualitative methods can be employed – recent literature on IE barely features qualitative methods, e.g. the World Bank's Independent Evaluation Group (IEG) report (2006) includes only one paragraph on qualitative methods, and the Center for Global Development (CGD) Evaluation Gap Working Group Report offers even less (CGD 2006). Moreover, while reference is frequently made to mixing methods, there seems to be a somewhat hierarchical view of what constitutes rigorous evidence, and how it is possible to assess impact. One commentator suggests that the rise of randomised control trials (RCTs) in development is part of a disenchantment with a lack of attention to rigorous use of participatory and qualitative methods (Prowse, 2007).

In response to this, many evaluators, academics and intermediaries have stressed that there are alternative approaches to IE (e.g. EES, 2007; Jones, 2009; Mackay and Horton, 2003). The counterfactual is just one among many types of causality, for which there are various alternatives (recognised in the natural and social sciences): 'generative' causality involves identifying underlying processes that lead to change (one method of this type uses qualitative methods to assess causality by understanding people's operative reasons for their actions or behaviour change (Bhola, 2000));[3] another approach takes a 'configurational' approach to causality, in which outcomes are seen to follow from the combination of a fruitful combination of attributes (Pawson, 2002). Many argue that is also possible to assess the counter-factual using non-experimental theory-driven methods, such as 'process tracing', which examines causation as part of a theory focusing on a sequence of causal steps. While these approaches are frequently dismissed by proponents of experimental IEs, the 'gold standard' perspective meets with considerable opposition. For example, evaluation bodies have spoken out on the need for a methodologically diverse approach to impact evaluation in development (EES, 2007).

These debates have been echoed elsewhere in the past. For example, RCTs were first used to evaluate public policy in the 1930s, building on earlier use of controlled experiments in psychology and education and drawing on scepticism about the ability of existing methods to firmly establish causation. They then blossomed from the 1960s in the 'great societies' programmes, but they failed to fulfil their promise and generated decades of fierce, polarised debates (Oakley, 2000). Therefore, possibly the first lesson to be drawn from the American experience is that these methodological and epistemological debates are not easily settled. It is for this reason that we will focus the remainder of this report on more practical issues, and questions about how IE functions in the real world. Since many interpret 'IE' to rightly refer only to quantitative, experimental (or quasi-experimental) methods of evaluation (referred to from here onwards simply as 'experimental IE' as a shorthand), and since many more who do not necessarily subscribe to this view nonetheless recognise that it is the dominant type of IE carried out in the sector, it is worth spending some time focusing on it in this study.

2.1.2 What does experimental IE tell us?

This section discusses some generalised characteristics of the knowledge produced by experimental IE. This is intended not to function as a critique of the method, but rather to recognise that (like any other method) it has strengths and weaknesses, and the knowledge it produces is useful to answer some questions but not others.

- **'Does it work?' 'To what extent?'** Experimental IE focuses on assessing whether a particular project/programme/policy had an effect on different measured outcomes in a target group. This is about 'proving' an effect, testing a particular model of intervention and asking what happened as a result. It requires quantifiable measures of the outcomes of interest, and generates information about the effects on various outcomes in terms of an 'added' quantity of those outcomes. This can be contrasted with other types of questions that might focus on

[3] Indeed, proponents of this approach argue that it is the only way to understand 'causality' in the social sciences: this is because the 'reasons' people have for their behaviour are not reducible to physical models of causation exemplified by the counter-factual (Jones, 2006).

improving policy or practice (e.g. 'how could we do this better?', or 'what worked, and what did not?')

- **Counterfactual attribution:** The device of the counterfactual usually involves comparing what happens with what would have happened in the absence of the programme, in order to understand what effect can be attributed to it. This provides information about what differences in outcome are associated with the intervention occurring (or, sometimes, with different elements of an intervention), and with what level of variation (etc.) This is quite different from asking (for example) 'what combination of factors brought about this change?' or 'what causal processes lead to the change?'
- **'Unbiased':** The experimental design has the function of removing selection bias (also known as 'sample bias'), which is a distortion of the evidence, a pre-selection of samples that may preferentially include or exclude certain kinds of results. This has the effect of increasing the robustness of the knowledge produced. Selection bias is one among a number of types of bias that can affect the quality of scientific evidence. Another type of bias is 'observer bias', where the researcher has an expectation of what the result should be and consciously/unconsciously affects the behaviour of the subject or their measurement.[4] Another type of bias is 'observer bias', where the person being studied alters their behaviour owing to their awareness of being observed, or as a reaction to their observer (e.g. surveyed people may respond differently depending on the gender of the surveyor, or if they are from government).
- **External validity:** An experimental IE is good at assessing what went on in a particular situation, but it needs to be supplemented with other knowledge to understand how the results can be interpreted outside that specific context. It is important to understand other questions, such as to what extent those conditions of implementation hold elsewhere, what processes helped produce an outcome, how similar interventions fared in similar (and contrasting) contexts, etc.

2.1.3 When is it feasible?

Assumptions implicit in the methodology of experimental IE shape the types of interventions that are amenable to them. We now briefly outline some requirements for conducting experimental IEs (many of them interrelated) that stem from the nature of the method. An important caveat to make is that there is diversity within this group of methods, and a number of methodological advances have been made that each would require a revision or nuancing of these statements. However, these approaches on the methodological frontier may not be so relevant given our focus on the realities of IEs in development policy and practice, and especially the practicalities of carrying out not necessarily academically driven IE, in contexts with low capacity, etc.

Timeframe: Looking at the effects of an intervention (for any type of IE) is dictated by the timeframe over which the outcomes of interest might feasibly be expected to occur. This will vary from case to case depending on the unit of analysis (e.g. understanding the effect of a cash transfer might involve looking at outcomes over several years; understanding the effect of the number of years spent in education might require a longer timeframe). Also relevant is how far removed the beneficiaries (or those of interest to the study) are from the activities of the project (e.g. building capacity of local government may take a number of years to have an effect on the poor in the area), looking down the 'causal chain' from the project intervention to beneficiaries.

'Dosing' model: Interventions must be able to be modelled as delivering a discrete and homogenous output. The population must be split into clear 'haves' and 'have-nots', a group of beneficiaries that receives the same output and a group that is entirely unaffected by it. There is also an assumption that programmes remain static for periods of time between measurements, or at least free of additional programmatic interventions.

[4] It is this type of bias that is eliminated using 'double-blind' RCT trials, recognised as the 'gold standard' in medicine, from which experimental IE has been developed. Of course, it would be difficult to perform a double-blind experimental IE, so this method cannot claim to be as unbiased as those techniques in medicine.

Clear, quantifiable outcome: The outcomes of interest must be readily quantifiable. That is, reducible unproblematically by the observer to an amount of some certain quantity. This may be easy in some cases (e.g. test scores, birth weight) but more challenging in others (e.g. empowerment, social capital).

Predictable dimensions and spread: In order to perform an experimental IE, it will be important to foresee who will receive an intervention and who may be affected by it (in order to assign control groups, etc.) It will also be important to specify what data to collect, which entails deciding for what outcomes one expects to see differences. This is in contrast with more exploratory methods.

Plausible counterfactual: In order to carry out an experimental IE it is necessary to find data that represent a similar case as those receiving the intervention in all relevant variables except for the fact that they are not beneficiaries of the intervention under study. Finding a sufficiently comparable context might be harder for some interventions, for example where the unit being studied is fairly large and change over time is a result of a number of factors, etc.

Complex situations: Experimental IEs are most suitable where it makes sense to attribute change to one intervention. They become less relevant where change comes about from the interaction of multiple factors, the actions of many different actors, etc.

2.1.4 Hypotheses

In short, like all methodologies, IEs have certain strengths and weaknesses, areas that they can usefully illuminate and others they cannot and certain inbuilt assumptions and epistemology. Because of this, they are suitable and useful in some situations and not in others.

This is relevant in the context of it being treated as a 'gold standard' for impact evaluation. This is like saying that a hammer is the 'gold standard' of tools: that is only the case if all your problems are shaped like nails. It is hoped that this section will build awareness of when and where experimental IEs should be used, and where it is wise to use other methods. The following hypotheses can be generated from this section:

1. *IE (of all types) is relevant only relative to the timescale over which an intervention might plausibly affect beneficiaries, so may require long timeframes.*
2. *Experimental IEs are most suited to interventions that have short and relatively simple impact pathways.*
3. *Experimental IEs are most suitable where an intervention can be modelled as involving discrete, homogenous outputs.*
4. *Experimental IEs require the intended effects of an intervention to be quantifiable.*
5. *Experimental IEs are only feasible in contexts where it makes sense to investigate what would have happened in the absence of the intervention.*
6. *Experimental IEs are suitable where effects are attributable to distinct forces/actions/interventions.*
7. *Where suitable, experimental IEs are able to provide robust evidence proving (and quantifying) the effectiveness of a project/programme/policy against predefined goals.*
8. *Experimental IEs are most suited to testing the effectiveness of a small number of interventions.*

2.2 Supply and demand: Commissioning, production and delivery of IEs

This section looks at what drives the production of IEs. With only limited published literature specifically addressing this question, we supplement the analysis with a discussion of what practical constraints there are to undertaking them, and for what purpose they might be commissioned. From this, we draw out tentative hypotheses related to who will commission them, on what sorts of project, where and when. We then compare this with a mapping of databases of experimental IEs to obtain some initial indications of how plausible the hypotheses are.

2.2.1 Practical considerations and constraints

A number of practical issues need considering when carrying out experimental IEs. Again, these are mostly well documented elsewhere, but it is worth highlighting the issues. While they do not apply all the time, and there are ways around each of these difficulties in different contexts, they constitute important considerations for actors seeking to understand the dynamics of IE production.

Capacity: Undertaking IEs requires a high level of scientific and professional expertise. However, this degree of technical sophistication is 'often lacking' in the field of applied development (Levine and Savedoff, 2006). This is especially so in the South, where low understanding of scientific methods is visible in government, civil society, non-governmental organisations (NGOs) and the general public (Jones et al., 2008). Rather, the expertise required to carry out experimental IEs exists at present only among a limited number of institutions and an army of independent consultants, largely in Northern contexts (Bloom, 2006; Foresti, 2007). This can, however, be mitigated by working in partnership.

Data: Ideally, experimental IE requires a comprehensive baseline survey, an end of project survey and, if possible, a mid-project survey. In addition, data need to be collected about households or communities that did not receive the intervention. This may be difficult, as there are often no baseline data available (Johnston, 2006; Prowse, 2007); where they do exist, data can often be poor quality monitoring data. Moreover, collecting data on groups not targeted by an intervention may be practically or politically difficult in some situations. There are a number of ways in which experimental and quasi-experimental methods can adjust to these difficulties. However, what this does mean is that it can often be difficult to make confident statements about impact, and rigorous studies may be overly costly.

Resources: Data and capacity difficulties in turn mean that robust IEs require substantial resources. This level of resourcing may be difficult to mobilise, particularly in developing country contexts.

Timing of supply and demand: Some argue that the evidence provided by IEs is most useful and relevant when designing a new project, and when seeking funding for a project (CGD, 2006). In order to carry out robust experimental IEs, certain requirements must be built into the project design, as well as the data collection activities. These two facts are in tension with the fact that it is rare for an evaluation to be the major concern at the outset of a project, and there are few incentives to undertake such evaluations. In addition, the (more interesting or robust?) results of such IEs become available only after programmes are completed, which is not the most useful approach for improving practice in a particular context (Foresti, 2007). To compensate for this, mid-term evaluations can be carried out, but they face two problems: the additional resources required and the fact that (depending on the particular context) impact may not be visible on shorter short timescales.

Ethical issues: There are, of course, ethical 'upsides' to IE. For example, the knowledge generated by the studies may be valuable as something that more effectively targets help at those in need. However, it is important to recognise that there are some potential ethical 'downsides' to experimental IE, which policymakers are likely to take seriously (Bamberger, 2008). At the simplest level, random distribution of interventions means deliberately denying it to some people who may need it, which may be troubling. A counterargument is that resources and interventions are inevitably limited and 'random' placement may be seen as 'fair'. This problem can also be compensated for through 'pipeline' methods, where the 'control' group receives the intervention in a subsequent phase. However, there are still some potential worries: some types of goods and services should be distributed based on need (Jones et al., 2008), and even within target populations there will be heterogeneous need – this means that random placement will conflict with the most equitable distribution to some extent. Moreover, in some contexts, the sheer urgency of problems may imply for some that minimal resources should be directed towards evaluation, in particular towards sending evaluators to study those in need who are not receiving help. There are also ambiguities: in the absence of reliable knowledge about what effects an intervention might have, or of levels of need, decisions are even more difficult. All of these considerations must be weighed up: the value of those who may suffer from unintended negative

effects of an intervention, and the potential unfairness of the distribution, must be compared with the long-term value of the knowledge that will be obtained for decision making.

Sensitivity to judgement: Because experimental IEs can go both ways – demonstrate positive or negative impact – an organisation that conducts them runs the risk of findings that could embarrass individuals, projects or programmes, and could undercut its ability to raise funds (Levine and Savedoff, 2006). There are many such disincentives to finding out 'bad news' existing in various organisations, and this is likely to inhibit the commissioning and/or publishing of many studies (Ravallion, 2005).

2.2.2 Potential purposes of undertaking IEs: Accountability and learning

Understanding how experimental IE suits the different purposes for which an organisation might commission or undertake an IE is another way to understand demand and supply. Bird (2002) identifies three main objectives of IE: improving practice (lesson learning), upward accountability and downward accountability.

Lesson learning: Where lesson learning is the priority, emphasis is often on achieving 'buy-in' from stakeholders. The focus is often on processes and exploring why effects occur, and timeliness is a particularly important factor (DAC, 2001). There are some reasons to suspect that experimental IE may not be ideal for directly improving practice in specific contexts: the focus on proving a quantified effect does not look at why an effect occurred, or the processes leading from the intervention to impacts. Trust and buy-in may also present a challenge given the 'detached' nature of the analysis and the capacity requirements, meaning that external consultants may be drafted in to carry out the evaluation. Without a mid-term evaluation, the timing of the evaluation results could also hinder lesson learning about ongoing programmes, as the results become available. However, such impact evaluations are likely to become more relevant for improving development interventions over the long term, as a large base of knowledge is built.

Upward accountability: Where accountability to donors is the priority, the virtues of rigour, independence and efficiency are prized. Experimental IE fits these requirements very well, providing robust and independent evidence 'proving' the effects of the intervention. It is also attractive to donors to commission evaluations that will generate results that are aligned with their wider goals, which are often system-level and beneficiary-level goals, such as poverty reduction or the Millennium Development Goals (MDGs) (White, 2005). As experimental IEs often attempt to detect effects on final welfare outcomes of beneficiaries, they are attractive in this respect. On the other hand, some commentators argue that it is inappropriate in complex situations to hold projects/programmes to account for 'impacts' that may not be feasibly predictable, and that an individual programme may only have a limited amount of influence over (Earl et al., 2001; Jones et al., 2008).

Downward accountability: Accountability to beneficiaries requires a participatory approach, allowing the questions to be defined by beneficiaries, understanding their views of the positive and negative effects of an intervention and taking into account their information needs. Experimental IEs do not seem at all well suited to this purpose for a number of reasons. For example, they are very unlikely to possess the technical capacities for undertaking them, and an appreciation of 'scientific' methods and rigour may be quite removed from local cultures (Jones et al., 2008).

2.2.3 Preliminary evidence

There is limited documented evidence on the forces driving the production of experimental IEs. However, some initial indications can be drawn from the above considerations, the available literature, the interviews carried out for our scoping with IE experts and from a mapping of databases of experimental IEs.

The available evidence does seem to point towards experimental IEs being commissioned in order to fulfil accountability purposes. Experiences over the past decade in CGIAR show that the primary function of experimental IE is to function as an accountability mechanism (Kelley et al, 2008). This is

backed up by the extremely high level of 'successful' experimental IEs being published. Watts et al. (2007) argue that many in policy and practice have become sceptical of the consistently high rates of return commonly reported in experimental IEs, and our interviewees concurred that negative results were seldom published. This is backed up in our database, which included nearly no evaluations that demonstrated 'no impact' or negative effects. This would also affect the lesson-learning potential of experimental IEs, as it is just as important to learn from failures as successes.

A common theme among our interviewees was that such evaluations are driven by political pressure and donors in order to prove effectiveness. Some see this 'exponential growth' in experimental IEs to be driven by powerful donors, and the World Bank in particular. Experimental IEs may be being advocated by these actors as the 'gold standard' to the exclusion of other methods, and one experienced evaluator suggested that many developing country policymakers have limited knowledge of other types of evaluation aside from experimental IEs. The low capacity to deal with experimental IEs is observed in developing countries (CGD, 2006), and backed up by a recent study on the perceptions of scientific methods and advice in policy in developing countries (Jones et al., 2008). Interviewees described evaluators as frequently being external Northern academics. In addition, it was felt that experimental IE is a 'method in search of an application', with the practical difficulties of carrying it out receiving disproportionate consideration. They tend to be carried out more where convenient than necessarily where needed.

We can therefore draw the following hypotheses about: Where have IEs been carried out? By whom? Which actors? In what sectors, what sort of projects? With what methods? Methods decided according to problems or vice versa?

9. *There are a number of potential practical issues in carrying out IEs, which (while each can be surmounted) nonetheless affects the when, where, how and by whom they are produced. Owing to these issues and the perceptions of those commissioning IEs, methodological concerns often receive disproportionate weight in deciding what and where to evaluate.*
10. *Experimental IEs are more likely to be carried out when they are expected to generate positive results. Like other types of evaluation, they tend to be published only if they demonstrate positive results.*
11. *Because IEs are still relatively new outside the health and agricultural fields, they tend to be undertaken on the basis of researcher or donor agency suggestion, rather than being demand driven.*
12. *The production of experimental IEs is driven largely by upward accountability to donors.*
13. *Experimental IEs tend to be commissioned less frequently to fulfil downward accountability, or operational learning purposes.*

2.3 Use and influence of IEs

This section looks at how the results of experimental IEs have been used, and what influence this has had on policy and practice. We discuss some of the factors behind the use (or non-use) of the results of experimental IEs, in order to suggest why they have been used in some instances and not in others.

An important caveat must be made here: there is not a great deal of analysis in the literature of how IEs are used in the field of international development. This is likely to follow in part from the underuse of IEs (CGD, 2006; Levine and Savedoff, 2006), and could owe in part to the lack of IEs undertaken, or lack of skills and capacities. It should be noted that this is indicative of a wider phenomenon of limited evidence on the use of evaluations in development (Sandison, 2005) and not specific to this type of evaluation. However, we argue that there are likely to be particular dynamics of use/non-use particular to the nature of experimental IE, and we will attempt to analyse these. In order to supplement the available literature from international development we will draw on evidence from outside development and theory of evaluation use and influence in general, as well as our expert informants again.

2.3.1 Conceptualisations of use

Table 1 outlines various ways in which evaluation use has been conceptualised. There is some degree of overlap between the different frameworks. For example, Sandison's instrumental use is similar to Patton's rendering judgments, whereas Sandison's conceptual use is akin to Marra's enlightenment use.

Table 1: Conceptualisations of evaluation use

Author	Types of use	Elaboration
Sandison (2005)	Instrumental use	Involves direct implementation of findings and recommendations to, for example, i) help decide whether to continue or terminate particular policy initiatives; ii) expand and institutionalise successful programmes and policies and cut back unsuccessful ones; and iii) figure out which programmes to modify and which components of the programme were in need of modification
	Conceptual use	Involves evaluations trickling down into the organisation in the form of new ideas and concepts – creating debate and dialogue, generating increased clarity and new solutions in the longer run (van de Putte, 2001), also providing a catalyst for change
	Process use (learning)	Involves learning on the part of the people and management involved in the evaluation
	Legitimising use	Corroborates a decision or understanding that the organisation already holds providing an independent reference
	Ritual use	Where evaluations serve a purely symbolic purpose, representing a desirable organisational quality such as accountability
	Misuse	Involves the suppressing, subverting, misrepresenting or distorting of findings for political reasons or personal advantage
	Non-use	Is where the evaluation is ignored because users find little or no value in the findings, are not aware, or the context has changed dramatically
Patton (1975)	Rendering judgements	Underpinned by accountability perspective (summative evaluation, accountability, audits, quality control, cost benefit decisions, decide a programme's future, accreditation/licensing)
	Facilitating improvements	Underpinned by the developmental perspective (formative evaluation, identify strengths and weaknesses, continuous improvement, quality enhancement, being a learning organisation, manage more effectively, adapt a model locally)
	Generating knowledge	Underpinned from the knowledge perspective of academic values (generalisations about effectiveness, extrapolate principles about what works, theory building, synthesise patterns across programmes, scholarly publishing, policymaking)
Marra (2000)	Instrumental	Decision makers have clear goals, seek direct attainment of these goals and have access to relevant information
	Enlightenment	Users base their decisions on a gradual accumulation and synthesis of information
Weiss (1999)	Direct	Occurs when information or findings are applied directly to change an action or alter a decision
	Indirect	Refers to a more intellectual and gradual process in which the decision maker is led to a more adequate appreciation of the problems addressed by the policy or programme
	Symbolic	This refers to situations where evaluation results are symbolic in that they are carried out simply to comply with administrative directives or to present an image of modernity

Drawing on these theoretical insights, we can outline three main conceptualisations of the ways in which experimental IEs should be put to use. These are not mutually exclusive: they involve clusters of ideas and assumptions, and are often implicit in texts on the method rather than explicitly stated. However, we argue that each represents an important force driving use. Once we have described the conceptualisation, we discuss how well suited to this purpose experimental IE might be.

Direct, instrumental use

One view is that experimental IEs should be used as a major input towards managing programmes based on results, at the operational level. They are seen to provide a major source of evidence to shape budget allocations among different activities (see e.g. Roche, 2000) based on robust evidence of 'what works'. They are also perceived as central to decisions to continue/discontinue/modify/scale up a smaller (possibly pilot) project (e.g. Duflo, 2004, argues that credible experimental IEs are required to ensure that the most effective programmes are scaled up at national and/or international levels).

This conceptualisation is particularly attractive to certain visions about decision making and the relationship between evidence and policy in development. A classical 'rational' model involves the decision maker generating possible solutions to a problem given her known goals, and then proceeds to analyse the costs and benefits of each course of action. Experimental IE would seem to provide an ideal input to this sort of process, as robust evidence of the benefits that an intervention will have in relation to overriding end goals. This is also related to a 'universal fix' view of the role of scientific research in development assistance, where a breakthrough can be replicated and applied with wide scope, and have a direct impact on poverty (Leach and Scoones, 2006).

At first glance, experimental IEs may be well suited to this conceptualisation, but there are potential problems.

- First, although experimental IEs may appear to give stark, clear evidence about whether an intervention is working (Orr, 1999), they tend to come with a number of important caveats, qualifications and nuances about what inferences can be drawn (often missed or glossed over by policymakers). Hence, they can be only one source of evidence to aid decision making, and will not provide robust evidence in and of themselves or as the major basis of a decision.
- Second, in order to be able to compare easily between different projects or courses of action one needs IEs of each option. This is in tension with the fact that (as we have argued) experimental IEs can be carried out to produce reliable results only in appropriate situations.
- Third, it is far from given that a pilot shown to be successful by an experimental IE can be easily 'scaled up' – Ravallion (2008) highlights a number of issues, for example the inputs may change (e.g. composition of who 'signs up' varies with scale), the intervention may change (resource effects) and the outcomes may also change (e.g. political economy effects, market responses).
- Finally, there is the issue of timing: since the effects of many programmes on end beneficiaries may in many cases take some time, the information may arrive too late to influence decisions over (for example) whether to scale up or terminate a project.

Legitimation

While it was initially seen as a type of misuse, legitimation has recently been recast as an acceptable application that contributes to incremental influence (Raitzer and Winkel, 2005). Many argue that experimental IEs function primarily to justify the actions of a particular organisation, project or programme. This 'symbolic' use serves to legitimise existing positions or decisions that have already been taken, and occurs particularly in the context of fundraising efforts.

In the context of questions about the effectiveness of development aid and pressures for funding, experimental IE seems very well suited to this. The highly credible nature of the knowledge provides an 'objective' stamp on the value of the programme. In the same vein, IEs produce information that can be used to construct a 'rate of return' for programmes (Kelley et al., 2008).

Although many factors suggest that experimental IE is well suited to legitimation, there are several ways in which it may struggle to fulfil this role:

- Depending on the type of project, if it has a long impact pathway it may be more of a stretch to robustly demonstrate impact on the welfare outcomes of intended beneficiaries. Other methodologies will be required in order to make rigorous judgements.

- While they may be used by organisations in managing their public image and staving off concerns about the effectiveness of development, this involves a particular assumption: that they provide the type of information most suitable to address the concerns of development sceptics. One interviewee suggested that, although experimental IEs are useful to highlight theoretical success (or failure) in a development intervention, public disputes about aid that motivate the development sceptics are instead about corruption, poor management and poor implementation.

Indirect use

The third perception about how experimental IEs contribute to policy and practice is in a more conceptual way, increasing the understanding of decision makers. One side of this is that they build up the stock of knowledge (viewed as a global public good – CGD 2006) about interventions that do or do not work in addressing particular policy and programmatic challenges. Another view is that they can create debate and dialogue, possibly refocusing debates or other shifts in the understanding of a particular programme or area of work.

This view involves recognising that the influence of experimental IEs tends to play out at more strategic and structural levels. The influence might occur sporadically, at critical junctures, and possibly where the IE is complemented by other types of information. Patton (1975) argues that the principal conceptual form of influence of evaluation involves a reduction in uncertainty for decision makers.

A number of dimensions determine how research influences policy in this way. The Overseas Development Institute (ODI) Research and Policy in Development (RAPID) framework looks at four dimensions that influence the bridging of research into policy: context (political and policy issues); evidence (quality and the way it is framed); links (between actors); and external factors (Court et al., 2005). This framework has been adapted to look specifically at the utilisation of evaluations (Sandison, 2005), where the key elements are quality, relational, organisational and external factors. Although many of these factors will vary from case to case, some features of IE and some general trends indicate that the following considerations are important:

- Contextual factors that affect utilisation include political dynamics, the need for organisations to protect their reputation for funding and external pressure for change. Teller (2008) suggests there is fear that negative evaluations play into the hands of the foreign aid critics among policymakers, which produces fear of the visibility of failures and mistakes. Patton (1975) suggests that evaluators should thus inform themselves of the political context of the evaluations on which they work. Through better awareness of the political implications and consequences of their research, evaluators can reduce their own uncertainty about the uses to which the evaluations will be put. Openness of the political system and a thriving evaluation community tend to make some countries more attuned to evaluation than others.
- Organisational culture also affects the utilisation of evaluation findings. Do organisations, for example, value learning and performance, and have accountability mechanisms and links to decision makers and knowledge management mechanisms? The sensitivity to objective evidence may hinder the commissioning of IEs, as organisations' will be concerned to protect their reputation, for funding and credibility purposes. Conversely, the clear value of successful IEs is likely to serve as a positive factor in securing future funding. Teller (2008) suggests there is common knowledge that many evaluations are not made public as they are procurement sensitive and then never released, too critical or poorly done. Patton (1975) showed that the negative or positive nature of the evaluation report was unimportant as a factor explaining utilisation, because findings in either direction were virtually never surprising. Surprises were more likely to increase than reduce uncertainty and good evaluation processes needed to build in feedback mechanisms to guarantee the relative predictability of the final report. Linked to this, studies that broke new ground were helpful as their potential for reducing uncertainty was greater, but they were viewed with some caution because decision makers clearly favoured the accumulation of as much information from as many sources as possible. Those studies that could be related to previous studies had a clear cumulative impact.

- Quality factors involve the design and planning of the evaluation, including the level of stakeholder participation, as well as the quality of the evidence (it should be credible and rigorous but also accessible and relevant). There should be mechanisms for follow-up and a credible evaluator. For IEs, potential problems seem to revolve around the timing of the results and immediate relevance to decision makers' needs, although Patton (1975), in his study of the utilisation of federal health evaluations, concludes that in no case was lateness in the release of evaluation findings or methodological quality considered a critical factor explaining either their utilisation or non-utilisation. Probably more important in the case of IEs is the perceived credibility of the approach and the potentially pragmatic nature of the conclusions such evidence can generate. The amount of resources devoted to a study may add to its credibility, but more costly evaluations have not shown any discernible patterns of utilisation different from less costly evaluations (ibid).
- Relational factors show the importance of personal links: trust, perceived credibility of the evaluator and a commonality of background and skills between evaluator and users can help; the evaluation unit and networks and communities of practice can serve to extend the links of key stakeholders, providing wider opportunities for influence (Sandison, 2005). The technical nature of IE and the tendency for it to be Northern driven may be obstacles. They have not tended to be carried out in a participatory way, and evaluators are likely to come from quite different backgrounds to those who use their work. However, where relationships of trust and links with key decision makers are fostered, experimental IEs may have significant influence on policy. Mechanisms for getting evidence into practice, as suggested by Davies (2004), include the integration of research understanding and use into the professional competences of decision makers and practitioners; getting policymakers and practitioners to 'own' the evidence needed to support and implement policy effectively; getting buy-in at the most appropriate levels; and getting policymakers and researchers to have a shared notion of evidence.

2.3.2 Evidence of use and influence of IEs in policy and practice

This section draws on available evidence to assess how the conceptualisations of use compare with the actual use of impact evaluations.

Direct

Available evidence from the development field shows that experimental IEs are rarely used instrumentally. Experience at CGIAR bears this out (Raitzer and Winkel, 2005), although important exceptions include the Progresa cash transfer programme, which has influenced the proliferation of social transfer programmes internationally (Behrman, 2007), and the uptake of school worming (Kremer, 2008). Similarly, Ruprah (2008) argues that, despite the production of a large number of rigorous evaluations by the Office of Evaluation and Oversight (OVE) of the Inter-American Development Bank (IADB), experimental IEs have not been adopted as the norm in the institution. Independent evaluation has not led to the identification and utilisation of lessons by the IADB, and hence the improved operational work that in turn leads to improvement in lives. This is not unusual – in general, direct influence is very rare (Sandison, 2005).

Lessons from other contexts bear out similar trends: Johnston (2006) also argued that political considerations (such as policymakers' desire to roll out programmes before evaluations were complete), other types of information and lack of confidence in evaluation results (relating to a perceived inconsistent use of particular indicators) have led to limited direct use of impact data in the UK, despite a general commitment to results-based policy. In developing and transition countries, those obstacles are also likely to be present, while the general environment may be less amenable to the evaluation process. For example, transition countries 'have been burdened with the need to pass legislation at a speed unknown in established western democracies … [and] that this creates time pressures that are at odds with optimal methods of policy and legislative development and decision making' (Ben-Gera, 2003 in Johnston, 2006).

In particular, it is worth reflecting on the use of experimental IEs for results-based management and in drives for effectiveness. A recent study from the US (Newcomer, 2008) examines such a system, known as the Programme Assessment and Rating Tool (PART), rolled out as a standard tool for performance across all government projects (including a stipulation for experimental or quasi-experimental IE). The study concludes that the initiative, driven by the office of management and budget, entirely failed in its purpose: it did not improve performance, with very little sign of use by any of the actors involved, as it was received as a drive for 'compliance' and 'legitimation' rather than being truly about evaluation and improving performance. Moreover, it was felt that the methods stipulated (experimental IEs) rarely represented the key measures of performance for agencies. Hopefully, these problems serve as a lesson on how not to institutionalise learning and accountability measurements in development.

Legitimation

Originally seen as a form of 'misuse', this has been recast as a legitimate application of findings in a manner that contributes to incremental influence (Raitzer and Winkel, 2005). There is a lot of evidence of legitimating and symbolic use of experimental IEs. For example, in CGIAR, it was felt by many donors that 'defence of budgets' was one of the most crucial roles it had to play (ibid). Some interviewees argued that it in fact dominates over other functions, used as a marketing device to prove the aid organisation's successful work to the general public. Transparency and legitimation are clearly conflicting objectives in all cases where actual development outcomes are not fully satisfactory (Michaelowa and Borrman, 2005). Other indications that legitimation may be a primary use can be inferred from the overwhelming prevalence of experimental IEs demonstrating success (Watts et al., 2007).

Indirect

Drawing on the CGIAR case again, it was found that experimental IE does play an important role, even though direct use of the findings of IEs may not be readily observable. There is evidence that this does go on, but it is harder to spot. Many argue that they influence decisions at the strategic and operational level (for example, Mackay and Horton, 2003). Balthasar and Rieder (2000) found that use of findings at these levels tended to be indirect and entails 'a complex process, which requires keeping in touch with decision makers at these levels and exploiting 'windows of opportunity when they occur'.

However, it is evident that there are a number of barriers to using experimental IE for 'learning' purposes, given other perceptions of use. For example, the focus on legitimation reduces learning potential through the observed publication bias.

Funding patterns and their impact on use

To many in the development community, the CGD paper 'When Will We Ever Learn?', which has helped spark renewed interest in experimental IEs, will seem to have a highly ironic title. If the gold standard continues to be adhere to by powerful donors, routinely requiring experimental IE for projects, this may have a detrimental or regressive affect on policy. If funds are influenced to a large extent by the ability to demonstrate impact using experimental IEs, it will lead to development policy and practice being skewed towards those types of projects most suitable to this methodology (and hence which find it easier to demonstrate impact). To many, the nature of projects that are particularly suitable to experimental IE contrasts starkly with the lessons learned over recent times about the complexity of social change (Ramalingam and Jones, 2008), and the importance of context (Leach and Scoones, 2006). Such contrasts include:

- Donor-driven and top-down as opposed to reflecting local needs, with intake according to local control.
- Project-based aid and technical/output-focused as opposed to multiple components; interventions responding to context; and the importance of advocacy, sector-wide approaches (SWAps), etc., for sustainable change.
- Demonstration of impact in the short term as opposed to a realistic view of the long-term nature of sustainable change.
- Emphasis on economic indicators as opposed to recognition of the multidimensional nature of poverty and change.

- Standardised, rigid project vs. frontline flexibility, evolving in response to changing conditions.
- Attributing effects to individual actors as opposed to a focus on working in partnership, harmonisation.

2.3.3 Coordination and collaboration

Coordination and collaboration efforts in the area of impact evaluation have been limited and difficult to achieve. Nevertheless, a number of coordination mechanisms have already been established in the field of impact evaluation. These include efforts at information sharing and dissemination (PREM/DIME, J-PAL, the UN Evaluation Group, the International Initiative for Impact Evaluation – 3IE, the Spanish Trust Fund for Impact Evaluation – SIEF and NONIE), capacity building (World Bank, J-PAL and SIEF), establishing a community of practice (UN), developing quality standards (World Bank and UN) and partnering with developing countries (World Bank, J-PAL, 3IE).

Two of the initiatives mentioned are elaborated briefly here: 3IE and SIEF. 3IE aims to provide and, summarize evidence of what works, when, why and for how much. 3IE reviews and synthesises existing evidence, updated as new evidence appears. 3IE also operates a grant programme, financing impact studies in low- and middle-income countries, and provides support to impact evaluation implementation.[5] SIEF supports the World Bank in evaluating the impact of innovative programmes to improve human welfare outcomes. SIEF supports prospective, rigorous evaluations in eligible developing countries, impact evaluation training, publications and dissemination of results.

The Campbell Collaboration from the social science field provides lessons for a more joined-up approach in designing and using impact evaluations. It aims to produce and disseminate systematic reviews of studies on the effectiveness of social and behavioural interventions for policymakers, practitioners and the public and follows a centralised collaboration model with a clearly defined shared vision. The Campbell Collaboration focuses on a limited number of audience-differentiated products (systematic reviews and meta-analyses, synopses of findings for end users and a web-based database of all studies reviewed). They also invest heavily in knowledge management. It follows a multi-layered organisational model that has a high degree of leadership in the form of a steering group that oversees general strategy and operations, and a secretariat responsible for coordination and dissemination activities. Networks in the humanitarian sector, such as the Active Learning Network for Accountability and Performance in Humanitarian Assistance (ALNAP), and the agricultural sector, such as Promoting Local Innovation in Ecologically Oriented Agriculture and NRM (PROLINNOVA), provide other effective models for research coordination and collaboration for impact evaluation use.

2.3.4 Hypotheses

It should be noted that a lot of these arguments are tentative in nature, and produced in order to serve as a starting point for analysing our six sectoral case studies.

14. *There are three main channels (not mutually exclusive) as to how experimental IEs can be put to use:*
    a. *Directly: Experimental IEs are a major input to managing programmes based on results. They can provide a major source of evidence to shape budget allocations among different activities and decisions to continue/discontinue/modify/scale up a smaller (possibly pilot) project.*
    b. *Legitimation: Experimental IEs are used to justify the actions of an organisation, particularly in the context of fundraising efforts.*
    c. *Indirect use: Experimental IEs contribute to policy and practice by building up the stock of knowledge about programmatic interventions that do or do not work in addressing particular policy and programmatic challenges. A conceptual way, creating debate and dialogue, generating increased clarity. This could be through strategic feedback, or through the knowledge generated.*

[5] See http://www.3ieimpact.org/page.php?pg=what for more details of what 3IE do.

15. *Of the different types of use, IEs are most frequently used for legitimation. This is largely in a 'defensive' mode, to protect funding. There is also growing indirect use.*
16. *Factors that affect/explain the use of IEs are:*
    a. *The rigour of experimental IE may be a strong force for its uptake.*
    b. *Relational factors such as trust and engagement between researchers and potential 'users' may be a large barrier to uptake.*
    c. *The fledgling state of IE knowledge management may be a significant barrier to uptake.*
17. *Where policy makers view experimental IE as the 'gold standard' and funding is influenced by the production of reliable experimental IE evidence, this risks skewing policy priorities towards areas most suitable for experimental IEs.*

3. Sectoral case studies

This section hones in on specific dynamics of IE production and use in different sectors in order to provide more nuanced insights about the suitability of impact evaluations for providing robust evidence on development effectiveness. Further, the extent to which the issues and hypotheses formulated in the previous section hold true is explored. Based on IE density in different sectors (from the annotated database, see below), consultations with the DFID Evaluation Department and other key stakeholders, six sectors were selected: health, social development (combined in the first sub-section below), renewable and natural resources, humanitarian, rural/urban development and infrastructure, with an additional case study looking at the production and use of IEs in results-based aid initiatives. Each sectoral case study was developed through between three and eight key informant interviews with both producers and users of IEs in both developed and developing countries – with almost 40 key informant interviews in the six studies undertaken. However, owing to a number of constraints, coverage was somewhat uneven, with a greater number of impact evaluation producers (such as donors and researchers) in the developed world (North America and Europe) than users, especially in developing country contexts. Subsequent work could usefully explore the views of the latter set of stakeholders.

The annotated database of IE studies compiled during the first scoping study (drawing on DIME, NONIE, PREM and Poverty Action Lab) was expanded to include IEs from the database of CGIAR. The database included information on thematic, sectoral and geographic areas, and on methodologies (quantitative, qualitative and mixed methods), followed by an analysis of coverage in terms of themes, sectors, geographic areas and methodologies in order to set the scene for more in-depth case studies. The database showed the largest proportion of impact evaluations have been carried out in the social development sector (41%). This is followed by agriculture/renewable natural resources (23%); private sector development/microfinance (10%); urban–rural development and infrastructure (8%); health (7%); other interventions (7%); and finally public sector management (5%). See Appendix 1 for more analysis of the expanded IE database.

Case 1: Human and social development[6]

Introduction

Commissioning, undertaking and using impact evaluations in the health sector, in particular, but also in the social development sector, have undergone rapid changes over the past decade. There have been considerable methodological advances, and a small but growing movement of researchers and supportive donors is seeking to push back frontiers of perceived methodological barriers. This culture of innovation has also played out in the involvement of stakeholders, especially Southern country governments and NGOs from both the North and South. Importantly, momentum has also been generated by a number of high profile success cases, where impact evaluations have been used to advance policy debates in some key pro-poor areas, including social protection and education.

Suitability and methodological approach

On a conceptual level, there is widespread agreement in the human and social development sector that randomised impact evaluations are appropriate when there is variability in coverage of an intervention in a population, whether this be geographical or categorical (e.g. poor vs. non-poor, eligible beneficiaries vs. non-beneficiaries). Similarly, there is general buy-in among our key informants that the gold standard of randomised quantitative impact evaluations provides strong statistical validity, but that it should be undertaken only if real life conditions permit. In particular, there appears to be a

[6] In addition to the 11 key informant interviews that shaped the analysis in this section, email correspondence was undertaken with Professor Jere Behrman of the University of Pennsylvania and Professor Edward Miguel, Centre for Evaluation for Global Action, University of California at Berkley.

healthy recognition that, while important, impact evaluations should be seen as only one component of broader efforts to institutionalise a stronger evaluation culture in the development arena. Political context variables, resource and capacity constraints and the likelihood of a receptive audience to findings all need to be considered and weighed carefully before undertaking such evaluations.

Interestingly, however, in contrast to oft-cited concerns about ethics and equity considerations (see discussion above), the randomisation debate in the human and social development sector appears to be increasingly framed as 'randomising is the ethical option'. In order to ensure that an intervention is creating more good than harm, the argument is that, wherever possible, robust evidence about the effectiveness of that intervention should be sought. It is inadequate to assume simply because the intervention is considered 'state of the art' that it is welfare enhancing and, in a world of scarce resources, not submitting interventions to the test of science may well be unethical (Bolton, interview 2008). This holds true even in the context of interventions for post-conflict traumas, as illustrated in Box 1. Moreover, randomising the process of programme rollout can help to overcome problems related to traditional patronage politics, where programme beneficiaries are determined more by political or logistical expedience than by equitable considerations.

Box 1: Criteria for randomisation in post-conflict mental health-related interventions

In order to assess whether interventions to improve the mental health of populations who have experienced conflict situations are really effective, the approach of the Center for Refugee and Disaster Response at Johns Hopkins University is to undertake impact evaluations so as to avoid attributing value to programmes that may simply be capturing the tendency for those who seek help to 'regress towards the mean'. In deciding whether or not to randomise treatments, the research team cover the following with the programme implementers, such as World Vision in their work with refugee camp populations in northern Uganda:

1. Is the evaluator convinced that the intervention will be effective?
2. The intervention cannot be too long – only weeks/ months – so as to minimise risk to both the control and treatment groups.
3. The evaluation team needs to ensure that groups with no intervention at first will not suffer permanent damage because they did not receive the intervention.
4. The evaluation team needs to monitor both the control and treatment groups to see that no one is in danger and to act if they should be deemed to be at risk.
5. The evaluation team needs to ensure that if the intervention is effective, that the control group will be the next to receive the treatment as the programme is rolled out.

Source: Bolton, interview 2008.

More broadly, however, there is considerable eagerness to move away from the 'bogeyman of randomisation', which unnecessarily polarises advocates calling for a more rigorous evaluation culture (Levine, interview 2008), and instead to focus on promoting more pluralistic approaches which help to maximise the strengths of impact evaluation and address its limitations (Bryce, interview 2008). More specifically, impact evaluations appear to be widely recognised as being limited in the sense that they are able to test only a small number of variables (Glennester, interview 2008)[7] and that they need to be complemented with process data to capture implementation dynamics and ascertain the extent to which the impact owes to the intervention alone or to the way in which it was implemented. Implementation issues are of course particularly pertinent in cases of weak or no impact. For instance, an education evaluation in Kenya found that textbooks had no discernible impact on children's scholastic performance; greater probing revealed that it was the level and quality of the textbooks that were lacking, especially in assisting weaker students, who were unable to read their contents (Glewwe, interview 2008).

Turning more specifically to the **health** sector, there has long been a consensus that impact evaluations have an important role to play in health, particularly in the area of medical interventions, such as vaccinations (e.g. Lu et al., 2006). This is in part because of the comparatively uncontested nature of the research questions involved and availability of measurable biomedical indicators. There

[7] Undertaking preliminary exploratory qualitative work helps to overcome this limitation to an extent by identifying promising variables to consider in quantitative impact evaluation work (Glennerster, interview 2008).

is a growing recognition, however, that in the public health field, where behavioural variables typically play a greater role, that impact evaluations of the RCT variety are of limited value in and of themselves. This is particularly because of the often complex causal pathways involved between interventions and beneficiary health outcomes (Victora et al., 2004). Instead, there is a call to combine a diversity of evaluation approaches in any assessment of impact, and especially to pay particular attention to process variables (Bryce et al., 2005a; 2005b). The Institute of International Programmes at the Johns Hopkins Bloomberg School of Public Health is therefore calling for the adoption of a 'stepwise evaluation approach', entailing the layering of evidence on the political and health sector context, service provision, utilisation and population coverage, impact and attribution of that impact to the programme intervention (Bryce, interview 2008). This approach is particularly important in areas of public health, such as child malnutrition or child mortality, where the current challenge is perceived to lie in scaling up interventions so as to increase population coverage, rather than testing the potential effectiveness of innovative interventions (Bryce et al., 2005a). Moreover, by focusing on effectiveness in real life developing country contexts, this evaluation methodology is seen as an important advance compared with a frequent reliance in the health sector on modelled results, which assume particular implementing conditions, which may not be realistic in diverse geographical, political and socio-cultural contexts (Bryce, interview 2008). Indeed, Teller (interview 2008) goes so far as to argue that the decision to undertake an impact evaluation should be based on strong indications that the intervention in question is innovative and effective and has the potential to be scaled up, and that, should effectiveness be confirmed through the evaluation findings, there are potential windows of opportunity in which the evaluation findings could be adopted at scale.

In the **social development** field, there is also a growing recognition that impact evaluations can be valuable in assessing the effectiveness of a range of interventions, from those designed to promote social capital, to those on human capital development and gender empowerment. This awareness is relatively recent, but has advanced rapidly over the past five years, with organisations such as the Poverty Action Lab, the International Food Policy Research Institute (IFPRI) and the World Bank playing a leading role in pioneering innovative methodological approaches to address these more contested social policy areas.

In the case of education policy and programming, this work has been partly shaped by a desire to move away from the limited rigour of regression analysis, and to develop more rigorous evidence on what interventions do and do not work – even if the scope of these questions is in the first instance less ambitious.

> In the 1980s and 1990s, empirical research was not characterised by the use of robust variables. The development economist crowd would carry out two million regressions – 10 studies would show positive results and 10 negative. As a result, policymakers ended up discounting all research as there was no quality filter. This shoddy research was annoying – it is better to answer small questions well than big questions badly (Muralidharan, interview 2008).

This shift in focus has been deemed particularly important in education given the high international and national priority attached to education sector investment in the MDGs and poverty reduction strategy papers (PRSPs). So research has focused on the relative effectiveness of educational software, such as flipcharts and textbooks, in improving children's scholastic outcomes (Glewwe, interview 2008), and interventions to tackle the widespread problem of teacher absenteeism (Muralidharan, interview 2008). The methodological approaches developed in Kenya (textbooks, flipcharts) and in India (teacher absenteeism) have since been adapted in other contexts, through general dissemination channels as well as involvement of the same personnel in similar studies in different contexts (see further discussion below). This knowledge transfer has been facilitated by the fact that there is a reasonable level of agreement on what the key policy challenges facing the education sector are in developing country contexts (Berlinski, interview 2008).

In the case of more contested social development interventions, debates about suitability have shifted considerably in the past five years, thanks to a willingness to combine quantitative and qualitative data collection and multidisciplinary analytical approaches in an iterative fashion in impact evaluations in

this field. In other words, there appear to be few proponents of solely quantitative impact evaluations among leading researchers and advocates of impact evaluations in the social development field. The Poverty Action Lab, for instance, has undertaken work on women's empowerment in South Asia, which began with up to a year of qualitative fieldwork to identify possible impact pathways between women's political empowerment – the introduction of reservation systems for women in local government (*panchayat*) leadership roles –and improved gender equality outcomes. By complementing quantitative evidence with these qualitative insights, researchers are better able to overcome the black box problem that plagues quantitative impact evaluations and thus to provide not only an assessment of the impact of an intervention but also insights into the mechanisms through which impact was achieved (Glennester, interview 2008). In the latter case, qualitative methods may be useful in identifying channels of both expected and unexpected impacts. Evaluators can undertake more in-depth discussions with either a representative sub-sample of programme beneficiaries or outliers of an observed trend in order to unpack the underlying dynamics at play. Such follow-up qualitative work can explore 'What has changed? What does the beneficiary ascribe to the intervention? What effects were expected or unexpected as a result of the intervention' (Bolton, interview 2008).

Importantly, such **mixed methods work** is also combining an array of qualitative research techniques: not only participatory research approaches (PRA) and tools, but also in-depth ethnographic observational work. As Adato (2008) argues in the case of the evaluation of the Turkish CCT programme, this is critical in order to balance how people articulate their motivations and behaviour with description of their actual actions. In some cases, anthropological techniques may be helpful, such as asking people to reflect on what the general populace does in a given circumstance (e.g. when a child is sick) rather than soliciting information on specific individual behaviour (Bolton, interview 2008), especially if the intervention touches on sensitive cultural issues (e.g. mental or physical health-seeking behaviour). In other contexts, it may be most helpful to triangulate what people say – which may suffer from a range of biases in the process of interacting with external researchers – with observed behaviour. It also helps to address the risk that, in some instances, participatory research may simply 'reinforce conventional wisdom' rather than elicit new insights (Glennester, interview 2008). For instance, impact evaluations on school attendance have identified that de-worming programmes have a significant effect on increased school attendance, as they help to address fatigue and diarrhoeal illnesses, but it is unlikely that focus group discussions with parents and children could have uncovered this. However, the impact evaluation commissioned by the World Bank in Kenya has subsequently played an important role in improving school enrolment rates in a highly cost effective way (Miguel and Kremer, 2002).

These innovative examples notwithstanding, a number of key informants recognised that impact evaluations could be strengthened if greater attention were paid to multidisciplinary design. While economists are increasingly turning to sector specialists, such as educationalists or gender experts, to better understand possible causal pathways, this appears to be on a largely *ad hoc* needs basis, rather than by forging multidisciplinary teams from the outset (Berlinski, Glewwe, interviews 2008).

Supply and demand

Overall, in the health and social development fields alike, impact evaluations remain more **supply driven**, but there is **rapidly increasing demand** from donor agencies, from some Southern governments, especially middle-income countries, and from a select group of NGOs and international NGOs. The dynamics in these sectors are quite distinct, however, reflecting the longer history of impact evaluation practice in the health sector and the greater level of consensus on appropriate interventions to address key health challenges in the developing world.

In the health sector, a key and distinct driver of the commissioning and production of impact evaluations is linked to **scaling up coverage of interventions**. A growing number of international health initiatives – from the International Health Partnership (IHP) to the Global Alliance for Vaccines and Immunisation (GAVI Alliance) to the Global Fund – are premised on the fact that many health problems in the developing world can be addressed by relatively simple and cost-effective

technologies (e.g. vaccines, bed-nets, presence of trained birth attendants, antiretrovirals (ARVs), etc.), yet millions continue to suffer from unacceptably high levels of morbidity and mortality owing to inequitable healthcare coverage. Scaling up considerations are therefore focused on expanding capacity and promoting sustainability over time. The implication for evaluation design is that 'in addition to addressing increases in coverage, evaluators should assess a broader range of factors that determine whether a program is likely to be successful, such as the quality of program design, the characteristics of the strategy or innovation, the interplay of various actors at country, regional and global levels the presence of champions and political commitment, organisational strength and the processes in place to learn from experience' (CI, 2008). This highlights the importance of the growing trend towards methodological pluralism and the combination of impact and process evidence in order to support policy decision-making processes.

A focus on coverage can be problematic, however, depending on the broader political context in which evaluations are commissioned and the underlying dynamics of the type of impact considered. For instance, given the focus of the US Agency for International Development (USAID) on diplomacy efforts in the context of the 'war on terror', the multi-million dollar President's Emergency Plan for AIDS Relief (PEPFAR) HIV/AIDS programme focused on ensuring access to ARVs among those infected with HIV rather than on prevention programmes and lives saved.

> The framing was that we needed a success story and in foreign policy terms this [PEPFAR] was successful … but what did they measure? Have you prevented new cases of HIV/AIDS? That should be the main objective. But instead the focus was on the number of new infected people received treatment – rather than determining whether the treatment was effective or reducing mortality … and at US$25 billion this should have been an important question … We have spoken out against this approach but the counterargument was that it was 'too early to document prevention'. 'It's an emergency programme'. This was the mentality (Teller, interview 2008).

In the social development sector, a primary driver of impact evaluations still appears to be **researcher interest**, which is then communicated and negotiated with international agencies (for funding) and typically only afterwards with implementing partners, whether they be NGOs or developing country governments (Glewwe, Berlinski, interviews 2008). For instance, the work on textbooks in Kenya stemmed from researcher interest to understand the relative importance of education 'software' in the promotion of universal enrolment (Glewwe, interview 2008), as did work on performance pay to address teacher absenteeism in India (Muralidharan, interview 2008). There is a general consensus that in this field too few programmes are undertaken on the basis of solid evidence, and that impact evaluators have a responsibility to breakdown the binary of doing vs. knowledge in development if resources are to be wisely allocated (Levine, interview 2008). More specifically, there is a recognition especially in the US context, that policy decision making in developed country contexts is typically informed by a much larger body of evidence than exists in the developing world, and that the pace at which robust impact evaluations are undertaken needs to be accelerated in order to accumulate a critical threshold of knowledge (Bolton, Glennester, interviews 2008). In the same vein, there is a push by researchers to undertake impact evaluations that are developed as such from the outset, rather than having to deal with less than ideal evaluation conditions of programmes that are already underway.

There does, however, appear to be **growing demand** for impact evaluations in the social development sector, as evidenced by the number of requests that impact evaluation specialists – both individuals and institutions such as the Poverty Action Lab – receive. In Mexico, demand is increasing on account of the creation of a dedicated national evaluation agency, the National Council for Evaluation of Social Development Policy (CONEVAL), responsible for assessing social programmes. Our stakeholder interviews suggest that this demand is not always well informed, with many requests coming from programme managers or funders, who want to evaluate impact at the end of the programme cycle but lack baseline and interim process data. Moreover, academics in the business of impact evaluation admit to being quite selective in terms of the types of questions and programmes for which they will agree to undertake impact evaluations. Many are interested in getting involved in evaluations not so much because of the potential to shape programme outcomes in a specific context, but because of the possibility of **contributing to broader knowledge about new interventions** (and appropriate research methodologies) in a particular field. As a result, there is a much **greater supply of impact evaluations**

on innovative programmes rather than replication studies to see whether an intervention is adaptable across diverse contexts (Glennester, interview 2008). Especially given the time constraints involved in providing evaluation findings so that they can feed into decision-making timeframes, the appeal of contributing to a global knowledge resource so that interventions have the potential of going to scale in a range of settings is a more powerful motivating force (ibid).

The importance of **learning** as a motivating force also appears to be an important factor in the involvement of a select number of international NGOs (such as World Vision, Save the Children and Oxfam) and Northern NGOs (e.g. Pratham in the US[8]) and Southern NGOs (e.g. Prenji Foundation in India[9], the Bangladesh Rural Advancement Committee – BRAC[10]) in impact evaluation work. As NGOs, especially Northern-based international NGOS, move increasingly away from a service delivery orientation towards policy advocacy work, evaluation findings can provide valuable evidence and also help to promote their re-branding as a learning organisation (Save the Children UK, 2009; Bolton and Ndogoni 2001). In the case of mental health, for instance, the partnership that World Vision forged with the Johns Hopkins Center for Refugee and Disaster Response in carrying out an impact evaluation of two key interventions (interpersonal psychotherapy and creative play) in northern Uganda for former youth combatants involved in the Lord's Resistance Army (see Bolton et al., 2007) has helped the organisation raise funding to undertake similar work in other contexts, on the basis that they have rigorous evidence to support the effectiveness of their approach (Bolton, interview 2008).

Box 2: NGOs and impact evaluations

Especially in the social development sector, NGOs play an important role in programme delivery in many development country contexts. In order to promote more rigorous evidence about the value of the work they undertake and also to increase the leverage of pilot projects (which is all that resource constraints typically allow NGOs to undertake) in policy dialogues, a number of NGOs have recently become involved in a number of pioneering impact evaluation partnerships. The following provides a brief snapshot of several of these initiatives.

Pratham is an Indian-based NGO focused on extending access to primary education. From origins in Mumbai in 1994, it is now present in 21 states in India. In 2000, Pratham was awarded the Global Development Network Award, sponsored by the World Bank and the Japan International Cooperation Agency (JICA) for its innovative work in promoting universal primary education. The organisation has been able to promote the effectiveness of its work through the use of low-cost participatory impact evaluation assessment tools, through a household survey on drivers and barriers to elementary education and the People's Audit of Health, Education and Livelihoods, which measures basic indicators through a combination of surveys, pictorial representations/ drawings, community activities and facility observations.

Azim Prenji Foundation, another education-focused NGO in India, has formed a partnership with the World Bank, the state government of Andhra Pradesh and the University of Berkley in implementing APRest, an impact evaluation on performance pay as an approach to tackle teacher absenteeism, especially in impoverished rural areas. Because of the NGO's perceived neutrality and the fact that NGOs typically have resources to work only in a select number of localities, randomisation has not met with any strong political opposition.

World Vision, an international NGO focused on poverty reduction, sustainable development and child wellbeing, has collaborated with a number of academics at the Center for Refugee and Disaster Response at Johns Hopkins Bloomberg School of Public Health to undergo an evaluation of its interventions focused on providing psychotherapy to post-conflict trauma victims in northern Uganda. This is part of the organisation's emphasis on research-informed policy advocacy. The evaluation findings, which have been peer reviewed in an international journal have in turn enabled the organisation to secure additional funding for its mental health work.

Sources: www.cgdev.org/doc/events/10.23.07/10.22.07/Paheli_oct22.pdf; Muralidharn and Sundararaman (2008); www.azimpremjifoundation.org/; and Bolton et al. (2007).

Despite this growing **NGO interest in impact evaluation as a learning tool**, it is far from universal. There appears to be 'a central tension in the development community between the value of doing and the value of generating knowledge' (Levine, interview 2008). While some researchers would like to

[8] See http://www.pratham.org/.
[9] See http://www.azimpremjifoundation.org/.
[10] See http://www.bracresearch.org/redupdates/Newsletter_Dec2007.pdf.

move towards a norm in programme intervention, whereby impact evaluation tools are embedded from the outset in order to promote greater and more rapid learning, the understanding and appeal of the benefit of such a shift appears to be limited at present (Glewwe, Bolton, interviews 2008).

Although not as frequently discussed by our key informants, **upward accountability** emerged as an important variable in shaping the demand for impact evaluations, especially on the part of funding agencies, such as the World Bank and the Spanish Aid Impact Evaluation Project,[11] as well as by a select number of Southern national governments (Ahmed, interview 2008). Accountability as a driver of the commissioning of impact evaluations has been particularly prominent in the case of the social protection field in Latin America. The demonstration effect of the Mexican CCT programme Progresa/Oportunidades has been powerful throughout the region (see Box 3 below), and feeds into broader high-level debates on governance, accountability and public performance in countries such as Chile, Colombia and Honduras (Rawlings, interview 2008). Interestingly, in this regard, World Bank Senior Economist for Human Development, Laura Rawlings, defined impact evaluation 'as a methodology with a counterfactual – with a control and comparison – drawing on reliable quantitative data plus cost measures'. In other words, cost effectiveness is seen as a critical dimension holding governments to account, not only for achieving impact but also for delivering value for money. Impact evaluations in turn feed into discussions on service delivery benchmarking and public expenditure tracking initiatives, which are about 'getting citizens accustomed to expectations that they can have of governments' (ibid). And for the World Bank itself, carrying out impact evaluations is a central component of its 'fiduciary responsibility'.

Perhaps more than any other developing country government, Mexico has **integrated the commissioning of impact evaluations with a concern with promoting accountability and transparency**. These efforts, spearheaded by the high profile and highly successful evaluation of Progresa, were institutionalised in 2006 in the establishment of CONEVAL, an institution charged with the monitoring and evaluation of the country's social development programmes. Impact evaluations are part of a broader effort to develop a rigorous and multi-pronged evaluation culture, which serves both accountability and learning purposes. The perceived importance of this culture shift is highlighted in the reliance on impact evaluation results by the current President (for a nutritional supplement programme) and a current state governor with presidential ambitions (for a housing project) in demonstrating their governing effectiveness to the populace. Major social programmes are required to undergo periodic impact evaluations and to publicise these results on their agency's website, along with an official response to the evaluation findings from the programme implementing agency and an action plan that is informed by at least some of the findings. This approach is designed to provide some space for implementing agencies to contextualise the findings within the programme's context realities, which they are likely to have a stronger appreciation of than external evaluators, and also to adopt the findings they find valuable rather than feeling compelled to take on board all the results in a top-down manner. The rationale behind the approach is that accountability (with a threat of 'punishment' for those found wanting) is in tension with the learning purpose of evaluations and there needs to be a way to marry the two to achieve both goals (Hernandez, interview 2008).

An additional **potential drawback of accountability** as a driver of evaluation that was recognised by our key informants was the potential for opponents of public expenditure on social services to be their greatest champion. Levine (interview 2008) pointed out that Republican Members of Congress have been more likely to commission or support impact evaluations on social programmes than their Democratic counterparts.

As discussed in Box 3, a critical reason for the effectiveness of Progresa was the **capacity** of key policymakers to understand the value of rigorous impact evaluations and the political will to use this evidence in decision making. However, in many other developing country contexts, a dearth of requisite

[11] See: http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:21419502~menuPK:384336~pagePK:148956~piPK:216618~theSitePK:384329,00.html.

capacities is no doubt a major barrier to greater demand for impact evaluations (e.g. Escobal, Galab, interviews 2008). There does, however, appear to be a growing movement among impact evaluation experts to undertake these in partnership with developing country researchers and/or government agencies. Glennester (interview 2008) of the Poverty Action Lab, for instance, insists that 'we can't do what we do without stakeholder engagement. I shudder at discussions of independence and "objective results". We have incredibly close relationships. We have researchers embedded in partner organisations full-time. We talk incessantly with the implementing partner.' Similarly, Bryce (interview 2008) emphasises that the Institute for International Programmes at Johns Hopkins works only through Southern researchers that have legitimacy in the eyes of the government, so that the findings will be more likely to be adopted. The work on teacher absenteeism undertaken by the World Bank in India has worked closely with state secretaries of education (Chaudhury et al., 2006; Muralidharan, interview 2008), and Berhman (2007) attributes significant weight to close links between IFPRI staff and Mexican Progresa programme managers in identifying the evaluation's successful outcome. In addition, the importance of impact evaluation training programmes run by the Poverty Action Lab and the World Bank were recognised as playing an important role in encouraging demand among Southern government officials for impact evaluations.

Communication and dissemination

There is a general agreement that, although important advances have been made in terms of knowledge sharing on impact evaluations, much more could and should be done. Overall, academic channels remain the most common dissemination and communication approach, i.e. publication of findings in peer-reviewed journals and at academic conferences. A number of respondents recognised the importance of networks and related websites, such as the Poverty Action Lab, the Yale Innovations for Poverty Action, the World Bank DIME initiative, NONIE, 3IE, the SIEF and the CGD's evaluation work, but as many respondents (especially academic researchers) were unaware of these specific initiatives. The role of the World Bank, the IADB and IFPRI in championing impact evaluations and promoting communication across countries and regions was also acknowledged as an important channel of knowledge sharing (e.g. Berhman, 2007; Ryan and Meng, 2004).

A number of important but still fledging initiatives also exist in terms of communicating evaluation findings to non-academic audiences. These include:

- The Poverty Action Lab's discussion paper series (e.g. Duflo et al., 2008 on teacher absenteeism) and policy briefcases, which provide an overview of what is known about interventions in a particular field, including cost data, in 'order to provide policymakers with a menu of options' to inform decision making.
- The World Bank's meta-analysis work on social protection mechanisms, including social funds (Chase, 2002; Newman et al., 2002; Paxson and Schady, 2002; Pradhan and Rawlings, 2002) and the forthcoming meta-analysis on impact evaluations on CCTs (Rawlings, interview 2008).
- The World Banks' recent poster presentation initiative on impact evaluations in the human development sector.[12]
- The Science Council Standing Panel on Impact Assessment's Science Council Brief series.
- The Centre of Evaluation for Global Action at the University of Berkley's initiative to post media reporting on the findings from impact evaluations.[13]

Nevertheless, there is a general recognition that ensuring that results are effectively communicated to policy audiences is largely the result of passionate issue champions, who are relentless in their communication of key findings (Levine, interview 2008). For most researchers, however, owing to increasing pressures to not only 'publish or perish' but also 'secure funding or perish', there is little

[12] http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTPROGRAMS/EXTPUBSERV/0,,contentMDK:21941145~menuPK:477927~pagePK:64168182~piPK:64168060~theSitePK:477916,00.html#IEpostersessions.

[13] CONEVAL in Mexico has also sought to engage with the media and has found mixed results. On the one hand, an initiative to provide capacity building for journalists on impact evaluations was met with very low attendance, primarily because the initiative was seen as not 'offering a story'. On the other hand, the media has had great interest in a new mechanism CONEVAL is piloting on a set of compromises from programmes on how they are going to improve their performance, owing to the recommendations done by external evaluations (Hernandez, interview 2008).

time to engage substantially with non-academic stakeholders (e.g. Glewwe, Bolton, interviews 2008). Funders will need to tackle these incentive structures by specifically funding communication and policy engagement work, rather than relying on this work to be done around the margins. Moreover, it appears to be important that funding is secured for a broader array of professionals – researchers are generally more motivated to broker their own research findings rather than those of others, so communication specialists could be important in order to ensure that the cumulative insights of multiple impact evaluations are effectively communicated to key policy decision makers and development practitioners (Glewwe, interview 2008).

Use and influence

In keeping with the broader literature on impact evaluations, there is little systematic analysis of how impact evaluations have been used in policy processes about human and social development and, in turn, the efficacy of such efforts. Some analysts argue that this is because there is a **dearth of feedback loops between implementing agencies and evaluators after the evaluation itself has been concluded**. While researchers are increasingly encouraged to demonstrate impact of their work by donors, there are few incentives for implementing agencies to communicate on how evaluation findings shaped decision-making processes; instead, researchers often get to hear about such effects 'informally, through second-hand channels'. For instance, Ahmed (interview 2008) notes that, although he was involved in a major evaluation of Turkey's cash transfer programme, which found that cash transfers had little impact at the primary school level but a major impact on secondary school attendance (especially for girls), he was unaware whether or not this had led to a shift in programme targeting by the Turkish authorities. Important exceptions to this trend are recent work by IFPRI tracing the factors that shaped policy decision making around food for education programmes in Bangladesh (Ryan and Meng, 2004) and the success of the Progresa evaluation in ensuring the programme's survival at a major transition junction in Mexico's political history (Behrman, 2007). Nevertheless, as we will see below, an important orientation of the impact evaluation movement in the human and social development sector is to focus more on generalisable policy lessons rather than the mechanics of specific programmes. This section therefore discusses a number of possible uses for impact evaluation findings, and then turns to a discussion of both successes and challenges in the sector to date.

The first broad type of use is **instrumental use**. Here, views were mixed on the extent to which impact evaluation findings could feed into shaping the implementation of specific programmes. Because of frequent turnover of key policy decision makers during the lifecycle of an impact evaluation, some argue that it is difficult to influence specific programmes as they unfold, especially because the findings are often delivered at the end of the project (Muraldharan, interview 2008). Others (e.g. Gonzalez, Rawlings, interviews 2008) argue that it is critical and a critical part of a learning agenda. Bryce (interview 2008) also emphasises that more pluralistic approaches to impact evaluations are critical to ensure timely and useable interim feedback. Building on this idea, Levine (interview 2008) proposes that funding for impact evaluations be reconceptualised in two intertwined tranches – the first from evaluation design through to evaluation findings, and the second where the programme plan is closely informed by the learning from the evaluation. This would help promote a model of evaluation where evaluation evidence is routinely integrated into programmes to refine and nuance practice.

The second type of use is related to **indirect or conceptual use**. A high premium is placed on this type of use by many supporters of IEs in the human and social development sector, especially insofar as impact evaluations help to distil out core components of successful programmes as a global public resource, which can be adapted to diverse political contexts. Here, CCTs are no doubt one of the best-known examples of impact evaluation-informed policy transfer across countries and regions. Other examples include the work on school enrolment and de-worming, work on teacher absenteeism, on textbook quality and on post-conflict mental health interventions. Indeed, respondents from the Poverty Action Lab and the World Bank emphasised that learning about use in the context of specific programmes was not a priority, and that contributing to the development of a more generalisable

evidence base and investing in meta-analyses of 'what works and what doesn't work' was a stronger motivating force (Rawlings, Glennester, interviews 2008).

Legitimising or ritual use is the third type of use. This type of use tends to play out at two levels – the selection of interventions to be evaluated and the uptake of particular results. The demonstration of effectiveness to stakeholders such as NGO funders or taxpaying citizens serves as a powerful incentive structure, creating a tendency not only to undertake evaluations of programmes thought to be effective but also to select and use evaluation findings that are positive. For instance, World Vision has widely communicated findings that interpersonal psychotherapy is effective in addressing maladaptive behaviour in post-conflict contexts, but has been considerably less vocal about the result that creative play approaches have little added value (Bolton, interview 2008). This was also the case with the Mexican Progresa CCTs and the mud-to-cement floor scheme designed to tackle sanitation concerns in rural Mexico: good news stories served the aims of politically ambitious programme managers (Hernandez, interview 2008). Conversely, negative results tend not to be published, in part because they are perceived to threaten the professional reputations and interests of programme managers and administrations. This is a reality of the power dynamics shaping stakeholder involvement in impact evaluations that researchers need to be cognisant of and take on board to shape communication strategies for evaluation findings (Teller, interview 2008).

The extent to which impact evaluation findings are influential in sectoral policy processes is shaped by a variety of factors (including political context, framing of findings, linkages between producers and consumers of new research, quality of evidence), which are in keeping with broader insights about the knowledge–policy interface (Court et al., 2005; Jones et al., 2008). First, in terms of **political context**, there was a strong consensus that an enabling political environment for the uptake of evaluation findings is critical. This was emphasised by the example of USAID under the Bush administration between 2000 and 2008, where results-based budgeting was prioritised at the expense of transparency and learning. The PEPFAR evaluation (Institute of Medicine, 2007) was an example not of evidence-informed policymaking but of policy-driven evidence seeking (Teller, interview 2008). The fact that the week after the 2008 election of Barack Obama a directive was issued within USAID to strengthen the agency's evaluation culture appears to reinforce this perspective.

The importance of **political will** is also demonstrated in the case of CCT programmes in Latin America. Although the Progresa evaluation is often used to illustrate the importance of quality evaluation evidence in shaping policy decision making, more careful analyses suggest that a considerable degree of political bargaining was necessary to persuade President Fox to retain the programme in 2000 (Behrman, 2007). Moreover, Hernandez (interview 2008) is careful to emphasise that evidence is just one factor among many in decision-making processes, which need to be recognised as inherently political. Political will at the highest levels is important, as is political will at the programme manager level – evaluations are likely to have an influence only if there is buy-in from managers, as their cooperation is critical in terms of obtaining quality data and integrating learning from evaluation findings into subsequent implementation strategies. This is also powerfully illustrated in the case of Nicaragua's CCT programme. Although evaluation findings were of a high quality and suggested that the programme was having a positive effect on human development outcomes, the CCT initiative was discontinued when the current Sandinista government came to power, as it was not in keeping with their broader political philosophy (Rawlings, interview 2008). Similar trends can be seen in the donor community. For instance, although the IFPRI evaluation of Bangladesh's food for education programme indicated that a positive relationship between food and school enrolment, donors and the Bangladeshi government chose to focus on problems identified by the study relating to the role of private sector providers to justify the discontinuation of the programme and a switch to a cash-based programme (Ryan and Meng, 2004). In this regard, a number of respondents emphasised the importance of undertaking political analysis in order to determine the feasibility of possible evaluation recommendations. In some contexts, the first best policy solution might lack political and/or administrative feasibility.

The power of political context notwithstanding, **quality evaluations are also deemed an important variable in determining influence.** Levine (interview 2008) argues that use is more likely if the data generated is perceived to be of high quality. Mediocre evaluations, where external consultants fly in for a limited time period and undertake a 'huge vacuuming up of documents and patch together a story from the numbers' is likely to lack impact. 'It may be the best available data but it's very bad and very inadequate to draw statements of causality' (ibid). The corollary to this is that technically sound impact evaluations appear to provide simple, clean policy messages that have fewer caveats and greater intuitive appeal. They are not perceived as 'lying with statistics or being overly complex'. Moreover, even if impact evaluations point to marginal improvements in the case of large social programmes such as cash transfer programmes in Latin America, which are now reaching millions of households and absorbing significant proportions of annual social expenditure allocations, a 15% gain in effectiveness constitutes a powerful message (ibid).

Box 3: Demonstrating effectiveness – lessons from Mexico's Progresa/Oportunidades

Strikingly, when asked about examples of influential impact evaluations, almost all respondents referred to the example of Progresa, an evaluation of Mexico's cash transfer programme, which aims at promoting children's (and especially girls') educational, nutritional and health attainment through regular monthly cash payments to mothers. The evaluation is known for its positive contribution to the survival of the programme in the context of a major political upheaval in Mexico's history (a shift in power from a long dominant ruling party to the opposition), as well as for its demonstration effect and providing inspiration for similar types of programmes around the globe, including in other parts of Latin America (Nicaragua, Ecuador, Brazil), in Turkey, in New York and increasingly in Africa (e.g. Ghana, Kenya, Malawi).

But what accounts for Progresa's influence? Document analysis and key informant interviews carried out by Behrman (2007) on behalf of IFPRI suggest that a range of factors were critical, from which we can learn important lessons. Moreover, these views were reinforced by our own stakeholder interview with a CONEVAL expert:

- **Academic rigour:** The involvement of world renowned academics in the IFPRI evaluation team helped to ensure that the findings of the evaluation were perceived as technically sound.
- **Evaluation team neutrality:** Considerable weight has also been accorded to the perceived neutrality of IFPRI as the evaluating agency. By virtue of being international, IFRPI was seen as being outside competing domestic political interests, and the agency was also perceived to carry 'less baggage' than the World Bank or IADB, following the association of structural adjustment programmes of the 1980s and 1990s with a long 'lost decade' in Latin America.
- **Good news sells:** The positive findings of the Progresa evaluation were also important in its uptake; it is unlikely that a negative evaluation would have had the same international resonance and appeal.
- **Programme manager buy-in:** The recognition by programme managers of the value of science in informing public policy and the importance of rigorous evaluations constituted a pivotal variable in the success of the evaluation process.
- **Policy engagement:** contrary to more romanticised perceptions, political bargaining was part of the equation, and political engagement based on the evaluation findings with President Fox and with key international players such as the IADB and World Bank was critical in shaping outcomes in the Mexican context.
- **Translation into local language:** Although a simple point, the fact that the findings were translated into Spanish helped to promote broader local buy-in in Mexican political circles.
- **Media enthusiasm:** In terms of promoting broad international awareness, the coverage of the story in respected international media sources, including The Economist and The New York Times, was also cited as an important factor.
- **Knowledge translation:** Active brokering of the Progresa success story by the donor community, especially the World Bank, IADB and IFPRI, has also played an important role in ensuring international uptake. These knowledge translation efforts have taken the form of personal communication with key officials as well as extensive citing of IFPRI's evaluation and other related research in core agency publications (see Behrman, 2007 for a detailed analysis). This role is deemed especially important given that policymakers themselves seldom cite sources for their decisions.

Source: Behrman (2007); Hernandez, interview 2008.

Ways forward

In terms of strategies to promote more strategic commissioning and uptake of impact evaluations in developing country contexts, a number of important recommendations emerged from the key informant

interviews. First, there was a strong emphasis on being cautious about the appropriateness of impact evaluations, with a recognition that it is neither possible nor desirable to have impact evaluations for all programmes. Impact evaluations need to be seen as part of efforts to institutionalise a stronger evaluation culture, and clear criteria should be advanced (including innovativeness, cost effectiveness, political and technical feasibility). Valuable lessons could be learned from the experiences of a number of Latin American governments, which appear to be at the forefront of such efforts, especially Mexico's CONEVAL initiative and similar evaluation efforts in Chile and Colombia. In Asia, the efforts of the Indian and Indonesian government were also highlighted as positive in this field. In the case of Africa, this was also deemed critical, given that operational issues of scaling up and inefficiencies in implementation under poverty, crisis and turnover are crucial (Teller, 2008).

In order to promote the use of impact evaluations for learning purposes, a number of suggestions were made. First, funders could play a pivotal role in reorienting the incentive structures that research and programme funding patterns promote and reinforce, especially in a world that is increasingly shaped by a 'funding or perish' edict. Funding could be structured to actively promote greater uptake of evaluation findings in subsequent programme rounds, and to ensure that evaluation is not approached as a *post hoc* addition to a programme but rather is a central component. Owing to funding and time pressures on researchers, more funding could also be explicitly directed to communication activities (including the employment of communication professionals), especially with non-academic audiences. Similarly, priority for funding could be given to researchers who work with Southern researchers and implementing agencies in order to promote capacities and ownership, as well as to evaluation teams that explicitly include a multidisciplinary perspective in their evaluation design and a mixed methods analytical approach. Here, several respondents emphasised that it is important to distinguish between countries, particularly middle-income countries, which already have strong evaluation expertise, and others, especially low-income countries, where capacities to undertake and critically appraise impact evaluations remain weak and underdeveloped. Investing in ongoing training programmes for officials and researchers was highlighted as an important approach which, to date, has been shown to have an important domino effect on the demand for a stronger evaluation culture.

Second, in order to promote a better marriage between impact and process evidence, efforts are needed to provide more systematic guidelines and even templates for the prospective documentation of contextual and programmatic factors (Bryce, interview 2008). This is arguably particularly important in light of the perceived need to promote replication evaluation studies – i.e. to adopt pioneering evaluation methods in new thematic areas to a variety of different settings in order to explore what successful programme elements do and do not transfer. For this, academic researchers are not necessarily required, and professional evaluation companies that would have different incentive structures could be considered.

Third, there was considerable support for clustering new evaluations undertaken in accordance with some commonly agreed on policy-relevant knowledge priorities in the sector (e.g. CI, 2008). This is seen to be in keeping with global poverty reduction initiatives, such as the MDGs and the Global Health Partnership, as well as the Paris Declaration principles of harmonisation and alignment. However, the birds-eye perspective of evaluation experts and donors needs to be informed by frequent feedback mechanisms with beneficiary populations at grassroots level in order to ensure that priorities are contributing to pro-poor and social inclusion objectives. Promoting communities of practice on a sectoral basis, such as the USAID-supported Measure Evaluation Project, which has evolved around health and nutrition issues, is also seen as a valuable channel through which to share information on methodological innovations (which are often highly sector specific) as well as evaluation findings.

Communication channels could also be strengthened by the introduction of databases where all impact evaluation proposals could be uploaded and findings only published on condition of compliance with such archiving procedures. This, along with possible anonymisation of information about implementing agencies, could provide important incentives to routinely publish negative as well as positive evaluation findings. A database which serves as a clearing house for new learning on human and social development sector programme interventions was viewed as particularly important, as was the

communication of research findings in a variety of formats, including the presentation of policy options with cost data.

Case 2: Agriculture and renewable natural resources

This case study assesses IE use in relation to projects, programmes and other interventions in the agriculture sector and the renewable natural resource sector. This includes work such as testing and distributing new crop varieties, work on livestock, natural resource management and forestry. The sector has quite a long history with experimental IE: this was a natural result of conducting RCTs (e.g. of seed varieties) to test crops, in areas focusing largely on technology-based interventions. As such, many organisations have wide and differing experiences and practices around them, but there is a general acknowledgement of the strong influence of the World Bank and CGIAR in promoting them.

Assessing impact depends on the gestation time of the intervention: Overall, our key informant interviews were of the view that it is always beneficial to try and understand the effects of an intervention, but that timescales are a key determinant of what is feasible. The sector must deal with extremely long timescales in some instances: for example, crop breeding might take eight years, then adoption by a local population might take that time over again. Areas such as forestry and natural resources management (NRM) work on even longer timescales. Seasonal and annual climactic variation can also obscure differences, especially in the shorter term.

Quantitative and experimental methods are applicable to some cases. Certain types of interventions can promote the use of quantitative and experimental methods. First, quantitative IEs are amenable to interventions with short-term effects, such as those that measure the impacts of the uptake of seed varieties. In forestry and NRM, this is rare, so it is difficult to carry out such quantitative IEs, especially in assessing final welfare outcomes. Second, they are suited to interventions that are well defined, narrow and discrete. Simple, technology-focused programmes are most amenable to this type of evaluation; even with these, it is still challenging to assess poverty impact 'further down the line'. A third area is interventions where 'dosing', i.e. exactly who receives the intervention, can be controlled. This is quite difficult with some interventions as, for example, people can share seeds with non-project targets and so 'contaminate' the control group. Finally, quantitative IEs can be more easily conducted on interventions with tangible, easily quantifiable effects, with well-defined success indicators. Hence, quantitative IEs are suitable for uptake, technology, productivity, efficiency and nutrition, but less so for activities such as agronomic practices (which are harder to observe) and NRM arrangements. In many areas of this sector, however, the question of whether a particular outcome is 'beneficial' may differ depending on the scale or level being looked at. For example, biodiversity may be beneficial on one level but not another. A number of outcomes of interest are also likely to be contested, for example in relation to receiving benefits of common resources.

Qualitative studies are suitable elsewhere and have their own benefits. Qualitative methods are good at looking at processes and complex phenomena. They can function to help determine causality, for example through pathway analyses. One argument is that qualitative studies are actually needed to establish causation – only then can we get at the reasoning behind behaviour changes. Qualitative work can dig into the 'differences' behind distributions, although they can still be susceptible to missing such factors (e.g. in some large-scale surveys), where carried out without proper attention to power and context.

Experimental and quantitative methods are suitable for only a minority of policy areas. These methods work, for example, on genetic improvements in crops and looking at the impact of uptake. They have less relevance for work with livestock and agronomic practices, and it is very difficult if not impossible to use them to assess institutional practices, policy research, NRM and institutional change. One key informant stated that 'no more than 25% of the sector is suitable for quantitative IEs'.

Demand and supply; commissioning and delivery

Background: There is a long history of evaluation in the sector. Traditionally, RCTs were carried out to test genetic improvements in crops, for example in research on seed varieties. It was then a natural extension to send agricultural economists along with the scientists to assess the effect on productivity and other such outcomes, of populations using outputs such as livestock medicines, seed varieties, fertiliser use and other technological interventions.

Push from donors, for accountability purposes: Given public concerns around development effectiveness, public relations departments of development agencies want to justify returns to investment, set against the context of long-term reduction in funding to the sector. There is thus more pressure to demonstrate 'results' and produce 'good new stories' as well as spurious quantitative information, much of which serves little to aid decision making.

Heterogeneity among donors: There is pressure from the World Bank (which is CGIAR's largest core donor) and other multilateral development banks for quantitative experimental IE, which has a huge influence on agricultural research and development. Bilateral donors, though, such as the Nordic plus group, seem to believe in methodological plurality, drawing on qualitative and quantitative methods in the context of innovation systems and partnerships.

Other actors: Evaluation by national governments is generally weak. Moreover, they are rarely consulted when IEs are conducted by development agencies, as the latter tend to conflate development effectiveness with aid effectiveness. Much involves working with private sector actors, who tend to prefer quantitative methods but focus not on poverty or development impacts but on factors such as profit, market surveys, etc. Some interviewees felt that NGO actors are weak when it comes to using quantitative experimental methods, perhaps owing to lack of sufficient capacity.

Experimental IEs are carried out where it is convenient. IEs tend to be carried out where it is easier or cheaper to do so in terms of a project and broader infrastructure. For example, they tend to be carried out in areas easily accessible from a project site or in areas which have adopted technologies owing to financial pressures. This can often lead to methodological problems, which can render the studies useless. Capacities for conducting IE are generally weak across the sector. In fact, several consultancies have sprung up to help produce them, with mixed results from outsourcing.

The qualitative and learning side is underrepresented: Demands for IE from programme managers tend to focus on learning drawing on qualitative studies. However, they may not necessarily have the resources for this. There is a lack of commissioning of substantive qualitative evaluations, especially those with careful sampling (not just 'pilot sites'). Moreover, many qualitative insights have not been followed up by quantitative studies, evidenced by (for example) the lack of any study on the gendered use of extension services since the 1980s.

Personal/organisation's cultural traits and tendencies matter*:* In donors, the type of method preferred has often been driven by the head of the evaluation agency concerned. Natural scientists have particular perspectives on research methods, but tend to have less understanding of the complexities of social change. They do believe that they are contributing to learning, despite being commissioned for accountability purposes.

Use and influence

Dissemination: Often, evaluations will be used only in the organisations that commissioned them. Critical voices around the effectiveness of different interventions are thus few and far between. Key informants emphasised that knowing whether interventions did not work was just as important for learning. This often limits use. Key informants also felt that policymakers needed brief products.

Direct use, allocating funding: There is a push by some actors to use impact evaluation as the major input to budgetary allocations, for processes of 'formula funding'. Key informants suggested that IEs

are hence frequently commissioned with a focus on the allocation resources within development agencies. However, they are rarely used this way: trialling and scaling up of 'widgets' tend to follow a fashion rather than being based on evidence. Programmes are much more frequently picked up based on the 'charisma' of the technology or policy model and the hype around an approach or particular policy narrative. This can often come when policymakers take study tours at successful sites but do not see the full picture of the project, then tending to make attempts to adopt the model without properly understanding the contextual factors that facilitated its success.

Defensive mode, legitimising use: Results are often used by development agencies to defend/justify their programmes. One key informant suggested that 'if a project/programme/organisation can demonstrate impact, then that creates a constituency for continued or additional investment'. There is also pressure to carry out IEs to estimate rates of return to investment. For example, the 2008 World Development Report on Agriculture for Development (World Bank, 2007) references a number of experimental IEs in order to demonstrate cost effectiveness. However, pressure to demonstrate results has often led to poor quality studies that are not useable.

Different actors: IE is often more persuasive to those not involved in a project. However, people working on it are more likely to want to see what worked and why – which is harder to do with quantitative IEs. Technocrats are driven by quantitative data, whereas 'their political overlords' are driven by human stories – they tend to want to put a 'human face' on issues.

Worry about policy implications of bias towards experimental IEs: As more complex, 'softer' types of intervention are carried out, which are less amenable to 'rigorous IE' using experimental and quasi-experimental methods, the requirement for 'hard evidence' and 'proof of concept' may be being used to discredit methods by some factions who are sceptical of 'softer' approaches to reduce funding for them (Horton, interview 2008). 'The relative ease of applying IE methodologies for crops research means that a stronger case is often made for investments going to crops than livestock or NRM. However, a lot more goes into the formulation of policy priorities than the results of IEs' (Freeman, interview 2008). 'More complex factors (often missed by quantitative IE), such as the gendered adoption of technology, are hence often ignored by policymakers' (Meinzen Dick, interview 2008). However, 'the effect is not clear' (Byerlee, interview 2008)

Ways forward

Robust decision making not RBM: Key informants suggested that the use of evaluation be used to support judgment and decision making, rather than to substitute for it: accountability goes well beyond just 'count-ability'. The philosophy of results-based management (RBM) has been adopted as a fashion in development, and has very negative consequences. A better way to work with IEs can be seen in the example of Embrapa in Brazil: performance measurement is carried out to decide on which centres need assistance – so a failure to demonstrate impact does not lead to cutting funding.

Change the level of IE: The responsibility of most evaluations should be placed back at the institute and programme level, where they can be designed and managed to support decision making, rather than serve primarily for external legitimation and accountability.

Change the focus: Evaluation priorities should be shaped around those of developing country governments. Appropriate methods should be based on the challenges faced. Moreover, IE must be based on wider reflection about what sort of evidence is required for timely and practical decisions. While information about impact is needed, key actors also need to know how and why these impacts are generated.

Improve communication and promote issue champions: There is a need to raise awareness among donors, and a need for funded organisations such as CGIAR to take a stand about the strengths and weaknesses, and hence suitability, of IEs in specific contexts.

Case 3: Humanitarian aid

While there are a number of established initiatives focusing on evaluations, and accountability and learning more generally (e.g. ALNAP), a focus on impact has recently emerged. This seems to be a result of the politicisation of humanitarian aid, the institutionalisation of the humanitarian sector and the changing nature of vulnerability in the environment, resulting in an increase in the number and scale of emergencies (Proudlock and Ramalingam, 2008). A number of sorts of methodologies are being explored, including RCT-type approaches as well as participatory methods (e.g. Catley et al. 2008). However, as of now, there has not been a great deal of evaluation of impact carried out, so it is too early to assess robustly the drivers and dynamics of production, use and influence.

When to evaluate impact: There is broad agreement that, since all projects ultimately aim to achieve positive impacts and minimise negative effects among beneficiaries, evaluating impact should always be considered. While 'impact' is generally considered a long-term phenomenon in development contexts, some argue that impact in the humanitarian sector can feasibly be achieved in relatively short timescales, for instance when the goal is to reduce mortality in the face of unfolding disasters. There is an open question of the extent to which impact should be assessed at different levels. Some interviewees concentrated on project-level evaluation, whereas some studies have focused on the performance of the system as a whole in the course of particular events. An ongoing initiative is looking to generate yearly assessments of effectiveness.

Methods

Many key informants emphasised the importance of considering context when making decisions about methodology. Further, conducting rigorous IEs does not necessarily equate to using quantitative methods. Proponents of quantitative methodologies argue that they establish causality and are able to tell us what interventions 'work' or not. A number of factors constrain the use of quantitative methods such as RCTs. First, quantitative studies necessarily assume a lot of information about the system being studied. This can be a constraint because unintended factors may be missed, as well as making it difficult to explore multiple models of change. Second, in highly fluid contexts, the relevance of knowledge produced by studies will quickly diminish, as contexts and political forces change rapidly. Third, on an ethical level, RCTs would not be conducted in genuine emergency response work (as this would mean denying people services). However, as rollout of services is typically slow, necessity would mean that randomisation could be possible.

Qualitative methodologies, on the other hand, are able to explore unexpected facets of impact, which tend to be a common occurrence in humanitarian interventions and can often end up representing key programme impacts. Further, while participatory IEs are seen as more suited to highly fluid, unstable contexts, these have rarely been conducted, owing partly to methodological challenges and partly to hierarchical and bureaucratic tendencies of humanitarian agencies. Recent work in this field seeks to address this gap (TUFTS, 2008).

What areas of policy should use what methodology to look at impact? Many felt there was a requirement on the work involving simple programme outputs (e.g. food, vaccines, blankets) and simple indicators (e.g. saving lives, nutrition, possibly more work in food programmes). This may not be possible, however: since quantitative methods are not appropriate in emergency response work, this leaves potential for using them in areas such as post-recovery (e.g. land tenure and livelihoods), disaster risk reduction or advocacy. Doubts can be cast even on some of these. For example, it is unlikely that quantitative studies will be relevant for advocacy, as this involves multiple actors and highly diffuse and uncertain effects. Further, there is a tension between 'need them where we do not know what works in a context' vs. 'cannot really use quantitative methodologies in an exploratory way'. Food programmes, for instance, might be too urgent for RCTs (on ethical grounds). In an emergency, it is agreed that qualitative work, and work which does not necessarily look at longer-term impact, should be carried out. We need to know things like: if people were reached, if the timing was good, if people were satisfied, the quality of goods and services and who was left out.

Supply and demand; commissioning and delivery; dissemination

Very few IEs have been carried out in humanitarian contexts. Impact is frequently included as a category to assess within wider evaluations, but little attention is ever paid to it. For example, the ALNAP database shows very few IEs, despite most study terms of references asking explicitly for this. In fact, ALNAP meta-analyses showed that impact is dealt with least satisfactorily.

Donors and accountability: Some argue it is too early to say what is driving the demand for IEs, because of the small number that have actually been carried out. However, there is a general impression that the recent surge of interest specifically in RCT-type approaches to IE is being driven by some donors, associated with their rising profile in development circles since the late 1990s – seen as a 'fad' by some. This is largely an accountability drive: pressure from the outside to demonstrate results. As such, they are often seen as more of a 'box ticking exercise'. These pressures are then passed on to other organisations to the extent that they are dependent on donor funding. However, donors are heterogeneous, so this is not always the case, and with most organisations they experience internal tensions. How different organisations have looked at impact has depended on particular agendas and particular people at particular times, and can flow between different positions as these factors/forces change.

Public image and 'good news stories': The pressure to carry out IEs has come from the need to demonstrate results – partly to improve public image and help with fundraising – as well as to account for donor investments. This has involved measuring outcomes against predefined objectives and marketing project successes strongly. This pressure is particularly common among international NGOs, which tend to be concerned about their public image and fundraising potential.

Less relevant purposes: While IE is often discussed as 'learning what works', IEs are often seen as part of a 'box ticking exercise'. Their usefulness is rarely a concern or driver of production. Accountability to beneficiaries is also seen as 'very low on the radar'. Reflective practice in general is quite weak, and does not promise to be well served by the emphasis on RCTs by some actors as the 'magic bullet'.

Difficulties: Owing to the need to respond to acute emergencies, and the longer-term nature of impact, questions of IE are often postponed. The ability to carry out IE comes down to resources – there seems to be a limited capacity for evaluations in general. Field staff have little interest or time for complex methodologies when they are busy saving lives. Some organisations work through networks of volunteers, who see it as much more important to focus on 'doing the basics right' rather than building capacity for sophisticated evaluation methodologies. Related to this, very few feel they have time to access evaluations.

Possible effects of these forces on policy and practice: Some see the push for RCTs as part of the (donor) wish for humanitarian work to be simple. They are picked up often as part of a push for results, and on RBM agendas. Some feel that these targets can in fact make it harder to build in reflective practice. For example, the need for 'good news' and pressure to legitimise funding often means that unexpected and/or negative effects are glossed over or entirely missed.

Use and influence

Worries about relevance: Many interviewees felt that the question of impact as it is currently framed is not the crucial problem. Clearly, it is crucial for the sector to learn from experience and pay attention to accountability, but it is not clear whether a focus on RCTs is the answer to this. In addition, in terms of 'good news stories', other factors are seen as more important: being seen to be at the scene of a crisis first, being seen to do something.

Using evaluations: The difficulties of using evaluations in the sector more generally have already been documented (Sandison, 2005). Hence, the selective relevance of different methods of IEs is unlikely to

result in similarly skewed policy, although this is also because donors tend to appreciate the complexity of change processes and do not use IEs as the sole basis for their decisions.

Ways forward

Incentives and a culture of learning: This includes monitoring, evaluation, impact and selecting appropriate methods. Awareness needs to be raised among key stakeholders, for example key targets within the ALNAP group include senior managers. IEs should be utilisation focused, promoting use among policymakers through early engagement. Moreover, IEs should be carried out through engagement between evaluators, agencies and local beneficiaries. It can be useful for people looking at impact to cast themselves as 'embedded researchers' rather than 'evaluators' (connotation of policing erodes trust).

Developing clear criteria for IE suitability: It is important to understand how agencies can make better use of IE methodologies for decision making, and to understand the relative value of different approaches in different contexts. One key area is for humanitarian agencies to be better at listening, and more informed about empowering local populations. Beneficiary perception surveys may be a pragmatic way of understanding effects of interventions and ensuring accountability. There is also a need for a system-wide evaluation, assessing collective impact, ongoing performance and including an assessment of the effects on beneficiaries.

Case 4: Rural/urban development and infrastructure sector

This case study assesses IE use in relation to projects and programmes in the rural/urban development and infrastructure sector. This is a broad category, comprising a wide range of different interventions including public and quasi-public utilities and facilities such as roads, bridges (transit infrastructure more broadly), sewers and sewer plants, water lines, power, communications (such as telephone, cell phones, and internet). This sector also comprises the built environment including towns, cities, houses, roads, buildings and other built infrastructure as well as other actions to improve the standard of living in non-urban neighbourhoods, countryside, and remote villages. While IEs in the sector are often seen as crucial, IE production and use are relatively new compared with, say, the health and social development sectors. Evaluation culture is thus shifting from a focus on outputs and outcomes to impacts, in parallel with IE methodological innovation.

Suitability and methods

Respondents generally agreed that IEs were one of a range of evaluation tools, suited only to particular types of projects or interventions in the sector. However, IE production was on the increase (as they were in the IADB and the World Bank) in line with the pressures to find interventions that 'worked', with donors and research organisations often actively seeking data and interventions to which quantitative IEs can be applied, rather than vice versa.

IEs produced tend to draw on quantitative methods: IEs, particularly in the infrastructure and rural/urban development sector, at the World Bank, IADB, IFPRI, the Chr. Michelsen Institute (CMI and the Japanese Bank for International Cooperation (JBIC) tend to be quantitative in nature, drawing on experimental and quasi-experimental techniques. In Vietnam, while most evaluations conducted tend to be qualitative in nature, the World Bank and the UN Development Program (UNDP) are increasingly using quantitative IEs to assess projects and programmes. At the IADB, for instance, the gold standard for IE is one which has a control group (to assess the counterfactual) and collects baseline and follow-up data, that is, a study which assess impacts before and after and with and without the intervention. Qualitative methods, although talked about, are rarely used to supplement the analysis.

Quantitative IEs do not provide all the answers: While quantitative IEs can give policymakers and researchers robust answers about whether or not a project is having the desired impact, they have a number of flaws. Randomisation tends to work well when the sample size of the unit of analysis (e.g.

households) is relatively large. Hence, these may not work so well when the sample size is small (and the unit of analysis is itself quite large, such as local municipalities or regions). A key challenge of quantitative impact evaluations is their inability to illustrate causal mechanisms. That is, they tell you if impact has been achieved, but not how and why – the so-called 'black box' problem. Lessons learnt can hence be very basic. The acquisition of qualitative data can address these challenges to varying extents. Such data may be collected through field visits, and enable impact evaluators to reconcile any contradictory data they may have found from different elements of a treatment or control group or explain unanswered questions (e.g. information and communication technologies – ICTs – and telephones).

Examples of answers that qualitative methods can provide: With water supply interventions, knowing how many people you have connected to potable water tells you little about how people use water. Qualitative data are thus required to address this gap. In Colombia, after the evaluation of an IADB-funded housing programme, quantitative outcome indicators showed improved welfare and satisfaction among beneficiaries. However, people started moving back to where they came from – areas considered relatively insecure. Qualitative data (from semi-structured interviews) later revealed that the new community lacked cohesion. Further, drawing on local knowledge through interviews, an IE found additional and unintended benefits of a World Bank-funded road building programme in Honduras on those who lived near the road itself. Qualitative methods in combination quantitative methods are hence better able to capture multiple and linked causes of different outcomes and impacts, as well as to explain how and why interventions are successful or not. They are also better able to capture the human dimensions of a project and its impact. Nevertheless, despite the rhetoric around the use of mixed methods in IEs, this is rarely translated into IE practice.

Matching interventions to IE: IEs are generally used to assess interventions to which they are most suited (which can generate some bias). For example, project designers at the IADB are usually asked whether their interventions are amenable to IEs (on technical grounds) before they are commissioned. If not, other evaluation tools may be used to assess their effectiveness. JBIC, for example, conducts IEs on interventions (chosen according to sector, size and location) for accountability purposes (although initially conducted mainly to improve practice, while also trialling new methodological techniques). Owing to their resource-intensive nature, JBIC tends not to conduct IEs on small projects, while counterfactuals are perceived difficult to construct for mega projects (see above).

Suitability of IE to different interventions: Within the infrastructure sector, IEs can be used to assess, for example, electrification; the best way of pricing, installing/extending telephone networks; installing and expanding the internet; or what content is most useful – health information, market price information or housing developments (where the unit of analysis is the household). Regarding social investment and regional development funds, the unit of analysis is the municipality. IEs are then conducted by comparing municipalities that received funds and those that did not. If programmes provided national level coverage, IEs can be conducted by assessing the relative difference in funds between municipalities or regions, and associated outcomes/impact. For example, in Chile, regional development funds were provided to all interventions, but fund levels in some regions differed from those in others. The control and treatment group was constructed accordingly to assess whether the additional level of funding/support made a significant difference. Hence, challenges in creating a counterfactual make IEs difficult to apply to these and other types of interventions.

Interventions where building a counterfactual, randomisation and hence removing bias are challenging: The impact of roads can be more difficult to assess using impact evaluations, as they can have unforeseen effects on prices and welfare outcomes. However, ex ante impact evaluations can be used to plan where and how to place roads initially. In electrification interventions (where randomisation is challenging), participants can be randomised according to whether an electricity line from the pole to the house is subsidised by the government or paid by the household. With large infrastructure projects, such as the extension to the New Delhi metro (with funding from JBIC), counterfactuals are very difficult to construct. Some argue that IE practitioners do not, and perhaps are not given sufficient resources to, use sufficient imagination to construct counterfactuals. At the other

end of the spectrum, when the intervention is small scale and tailor made (such as those funded by social investment funds, say), it is difficult to draw general lessons, owing to their specificity

Other key challenges include lack of resources: Donors commit inadequate time and money to undertaking thorough impact evaluations, which are time and financially intensive. For example, five to seven years may be needed before any observable outcomes/impact can be seen as a result of, say, a rural electrification programme. Limits on political terms mean policymakers can rarely factor such long time spans into their planning, and contractors are reluctant to reason with those who commission them owing to financial pressures. Moreover, impact evaluators may have trouble acquiring relevant data, as statistical units/offices may have few resources, unless impact evaluators fund the acquisition of special data (at additional cost).

Attribution in multi-donor initiatives: Increasingly, multiple donors are pooling their funds to support one or more interventions, partly to reduce transaction costs for donors and recipient governments and partly to share risk. This creates challenges when donors have to account for aid funds spent to ministers and tax payers in their home country. CMI faced this challenge when conducting an IE for the Norwegian government on a multi-donor-funded hydroelectric project in Mozambique (which they had been involved in for about 15 to 20 years). They simply decided that, since Norway provided half the financing, they would also be responsible for half the impact.

Supply and demand

The commissioning of IEs tends to be driven by bilateral and multilateral donors, often through conditions attached to loan and grant agreements. Production and use of IEs, mainly for accountability purposes, have been promoted by trends in the US, the World Bank and, to a lesser extent, the UK. Civil servants, say in the Norwegian Agency for Development Cooperation (Norad), are now more likely to follow their counterparts in more powerful institutions such as the World Bank and commission IEs to evaluate development interventions, such as the building of hydroelectric power stations in Mozambique.

Political pressures to conduct IEs: Further, ministers of overseas development and/or foreign affairs agencies in developed countries will demand quantitative IEs be produced (owing to their intuitive appeal and their economics background) to assess impact of key interventions to satisfy certain constituencies, even if those interventions may not be amenable to IE use. More focus on results than on process mean that those contracted to conduct IEs are rarely provided with sufficient time and money to conduct robust and meaningful IEs. Supply and demand dynamics are also affected by the relationship between the financial and technical elements of the development agency. For example, the separation of Norad from the Ministry of Foreign Affairs means that the former has less money at its own disposal and the latter is more of a 'political animal' and harder to bargain with for resources. Meanwhile, in Japan, JBIC (the financial arm of Japanese aid) and JICA (the technical arm) are merging, among other things to promote more coherence between what is demanded by civil servants and what is carried out by researchers.

Actors involved in driving the IE agenda: In the multilateral banks (such as the World Bank and IADB), production of IEs was generally driven by those high up in decision-making structures to assess ex post what impact their funds were having on infrastructure development projects. However, IEs are increasingly being factored into project design (ex ante) at operational levels. This has been driven by a growing pressure to provide evidence of what works and what does not, as well as advances in IE design making them more cost effective. However, within the infrastructure and rural/urban development sectors, the absence of an IE 'movement' (in both developed and developing countries) has slowed both IE production and use compared with other sectors. IE movements (comprising researchers, policymakers and practitioners) in sectors such as health and social development have, among other things, promoted the development of methodological tools (around building counterfactuals, for example), enabling application of IEs to increasingly complex interventions.

A high level of expertise is required to conduct IEs: For example, JBIC, to ensure credibility with partners and competitors, ensures that those who conduct IEs have PhDs. This often results in highly academic debate around IEs. The World Bank mission conducting an evaluation on rural road and poor area development in Vietnam comprised top-class technical researchers. Implications are that key messages from IE findings are often not understood by those not literate in such debates.

Few developing country governments attach high priority to evaluation, let alone to production and use of IEs. Fears that donors may withdraw their aid on dissemination of 'bad news', little understanding of the role that IEs can play in piloting interventions before scale-up, as well as lack of time and resources (IEs can be expensive: costs of generating data, staff training and using experimental methods) could explain partner governments' lack of engagement. IEs are also seen by policymakers in developing countries as too academic in nature and not pragmatic enough.

Donors do make some attempts to engage with partner governments around IEs. IFPRI goes as far as working with those from implementing agencies (e.g. transport ministries) to undertake assessments and building appropriate capacity of researchers. It also works with policymakers to build their understanding of IEs and hence promote their uptake. Norad conducts IEs jointly and often with acceptance of partner governments. The IADB and World Bank only go as far as informing partner governments that they intend to send a mission to conduct an IE and welcome collaboration. World Bank projects often require governments to engage (through regular progress meetings) on evaluation issues, as was the case with a recent road building project in Honduras. However, partner governments tend to provide relevant data rather than deploy/second staff to work with IE donor teams. In Vietnam, although evaluation is a growing area, government may be unable rather than unwilling to deploy relevant staff, owing to a lack of sufficient capacity to engage critically with IEs. On a related issue, the World Bank is investing significant amounts in building the capacity of the general statistical office there, to help monitor progress towards poverty reduction targets and provide data for IEs.

Many developing country governments are starting to think more critically about evaluations more broadly, as well as IEs specifically. In Chile, the government is increasingly demanding cost-benefit analyses of investments. In Vietnam, the National Assembly is asking for evidence of impact to inform policy decisions around scaling up coverage. JBIC's Evaluation Department has had dialogue around IE production with some developing country governments in Latin America, such as Ecuador. The governments of Chile, Mexico, Brazil and Colombia have established evaluation departments within key ministries, including those of finance and planning. Levels of capacity vary, however. For example, in Chile, within the Department for Evaluation (located within the Office of the Budget), there are good linkages between evaluation and budgeting. This is not the case in Colombia. Donors are investing more money in improving evaluation capacity in many developing countries. For example, the World Bank is building the capacity of the General Statistical Office in Vietnam – including funding staff time and data collection activities and linking them to networks to share learning and best practice.

Communication and dissemination

In-country, dissemination of IE findings tends to be weak. IE findings in the form of a report are usually sent to the relevant ministry, in many cases the transport ministry. Policymakers have little time to read IE findings in any detail, though, and take away only high-level messages, if there any. Donors such as the World Bank hope that IE (and other research and evaluation) findings can be disseminated and/or accessed more widely, to government and non-governmental practitioners as well as the public. This is more likely to happen, and is happening in middle-income countries (with key exceptions such as China), where there are trends towards greater openness, access to information (through, e.g., the internet) as well as pressures from external donors (even as they become less dependent on aid).

Dissemination within donors and donor country audiences: Within donors, findings are first communicated internally, particularly to programme designers. Once they have been signed off by top management, they are usually disseminated externally to a range of audiences using different media. JBIC-sponsored IE findings are communicated by JBIC staff to bureaucrats and politicians through

seminars and direct communications (meetings and telephone conversations), academics through meetings and journal articles and the public through the television, radio and internet. A key challenge is that the public tends not to understand the significance of findings, whereas policymakers think findings are difficult to apply, owing to their complex and academic nature. This may be a flaw of the communication strategy (and lack of a message development process) rather than of the data *per se*.

Communication between contractors and clients: In donor countries, once IE findings have been released to the client, there tends to be little communication, if any, between the researcher/analyst and client. Moreover, little is done to build the knowledge base in the sector. Clients such as Norad look at different IEs in isolation, rather than in relation to previous studies, owing to a lack of meta-analyses. These issues are exacerbated by the high turnover among civil servants.

Use and influence

The use of IE findings by partner governments is limited. They are often content to account for investments through receipt of IE reports. There are cases where IE findings have been followed by appropriate policy decisions, but the extent to which this was informed by the findings is difficult to gauge. In Chile, an IE on a scheme to promote house ownership showed that the poorest were not able to repay their housing loans, not because of moral hazard (since the government provided people with loans) but because of their poor status. The findings were released shortly before the Housing Ministry dropped the loan element to the poorest, who instead received grants. The Ministry was satisfied that the IE findings were in line with their decision, but said nothing about whether the findings informed their decision. In Vietnam, findings from a recent impact analysis on rural roads and poor area development (which generally showed the programme was working well) were sent to the Ministry of Transport, but were used by its co-sponsors – the World Bank and DFID – to justify continued funding.

Legitimising use in donor organisations: Within the IADB, IEs are often used at Board level, with high performance programmes selected to legitimise policy decisions. However, there is a policy that all IEs be published. If results show an intervention is performing poorly, there will be extensive discussion, usually involving the programme designer, with findings both verified (were the results produced in the right way) and validated (were the results right).

IEs with poor results: Rather than focusing on improving the intervention, the assumption is that IEs with poor results reflect an intervention that is bad for people. Policymakers may choose to ignore IEs with poor results, as this could lead to a suspension of aid. Conversely, policymakers tend to use IE findings if they show an intervention to have a positive impact.

IEs in relation to pilot studies: IEs are most useful when interventions are being piloted. This rarely takes place in developing countries. China undertakes a considerable number of pilots, with successes (assessed through IEs, among other tools) going up on a huge scale. Piloting can take one or more years, which creates challenges for policymakers with strict term limits. With huge amounts of aid and private finance (increasingly through compacts) due to go to Africa for infrastructure development in the next years, there are opportunities for IE production and use, especially on piloting of interventions.

IE findings vs. methods: Results are more important that methods (Tuodero, interview 2008). Policymakers find quantitative results more persuasive and perhaps more intuitive, using them to, for example, corroborate decisions. Nevertheless, qualitative methods may better capture the human dimensions of an intervention and its impact and thus be more useful for policymaking.

Case 5: Impact evaluations of results-based aid

This case study is more exploratory than the sector cases. This is because results-based aid is a relatively new form of aid. In some cases (e.g. cash on delivery), programmes are just starting to be implemented, and IEs are often still in the planning phase; in others (e.g. results-based budget support), programmes have been running for a few years but IE methodologies are still being

developed. Results-based aid was not investigated as a separate 'sector' in the initial scoping study. This case study is based on a review of key resources and interviews with experts.

What is results-based aid?

Results-based aid comprises a variety of initiatives of recent years, including performance-based aid, outcome-based aid, output-based aid, outcome-based conditionality and cash on delivery aid. They all have in common an explicit linkage of future aid disbursements to some measure of results. The target results as well as the money available are agreed mutually in advance. The key distinction of results-based aid is actually not the focus on results (all good aid programmes focus on results), but the explicit linkage between aid disbursement and results rather than the normal linkage to inputs.

Results-based aid has been promoted for two reasons. First, there has been an acknowledgement that policy-based conditionality has not worked in promoting lasting change. Successful policies have to be home grown, and there needs to be sufficient policy space for governments to experiment with policies most appropriate to national circumstances. Second, there is an increased recognition that development efforts should be focused more strongly on results than inputs. A greater focus on results should in turn allow for a greater degree of ownership and accountability in aid relationships.

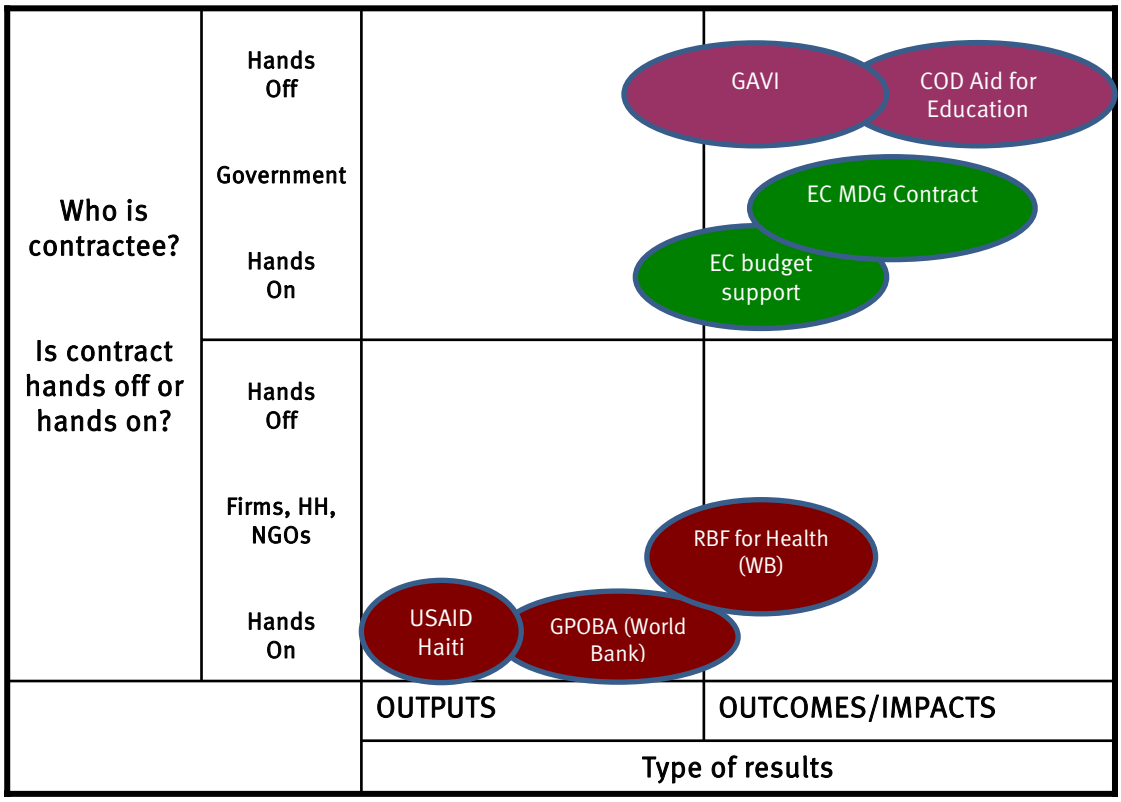
Results-based aid initiatives have some aspects in common but also differ on a number of dimensions:

- **Who is making the contract?** Results-based aid initiatives include agreements with government (macro) or with the private sector or NGOs (micro), or a combination of both.
- **Choice of results?** Aid disbursements are made dependent on outputs and/or outcomes. Some programmes involve a menu of results, whereas others specify a single output or outcome.[14]
- **Level of control?** Some results-based aid programmes have a 'hands off approach', focusing exclusively on whether or not the results are achieved and allowing delivery agents total freedom in how they are achieved. Other programmes are much more 'hands on' and will involve greater control on the part of the donor.

Figure 1 presents a number of results-based aid initiatives using the three dimensions described above.

[14] Other differences between different programmes with respect to results targets include i) whether results targets (e.g. payment per child educated) are set for individual countries or common across countries and ii) whether targets are fixed or proportional.

Figure 1: Results-based aid initiatives



Based on this figure we can distinguish three broad groups of results-based aid:

Output-based aid programmes (e.g. GPOBA, RBF for Health, USAID Haiti): Output-based aid initiatives consist of payments to providers (not to governments) that are linked to outputs (e.g. water connections). One of the earliest examples of this type of aid was a health programme in Haiti financed by USAID. In 1999, USAID developed a performance-based health service programme with NGOs in Haiti. Performance of NGOs was measured against a number of indicators, and payments were made against an output-based schedule. The initiative was further developed by the World Bank and led to the establishment of the Global Partnership for Output-based Aid (GPOBA). This is a multi-donor trust fund administered by the World Bank with a purpose to fund, demonstrate and document output-based aid approaches. So far, the fund has been focused largely on infrastructure but is now expanding into social sectors, such as health and education. More recently, another trust fund was established, on Results-based Financing (RBF) for Health.

Cash on delivery aid (e.g. education pilot, GAVI): Cash on delivery (COD) aid is a concept recently developed by the CGD (Barder, 2006). A number of donors have expressed an interest in piloting the approach in education. COD disburses aid according to units of progress, with an emphasis on outcomes rather than inputs. The outcomes have to be closely related to an objective that is shared by the donor and recipient and have to be measurable in a way that is continuous, making it possible to reward incremental progress. In contrast with most forms of results-based aid, COD aid requires recipients to assume full responsibility for the design and implementation of strategies to make progress. Donors do not specify or monitor inputs; rather, they contract independent verification of progress and pay, as agreed, for improved outcomes. The COD aid concept was developed based on experiences of the GAVI Immunisation Services Support programme (ISS). This performance-based strategy makes continued funding conditional on improved performance and high-quality coverage data. It allows countries and governments to spend ISS funds in any manner they deem appropriate; but funding is based on increases in the number of immunised children. Generally, countries are approved for five years of support, with the first year of funds (paid in instalments over three years) considered investment funds, and the subsequent four years 'reward' funds. The reward funding is

calculated at US$20 per additional child receiving DTP3 vaccinations above the number of children in the baseline year.

Results-based budget support (e.g. EC budget support)

The European Commission (EC) has introduced a new form of budget support in recent years, including a fixed and a variable tranche (EC, 2005). The amount of the variable tranche depends on whether the recipient meets mutually agreed targets on a range of public finance, health and education outcomes. Equally, the more recently launched MDG Contract is a form of budget support in which a minimum of 15% of aid is made dependent on results, in particular the MDGs.

Evaluations of results-based aid

Evaluations of results-based aid have tried to answer three broad types of questions:

- To what extent approaches and frameworks were in line with original objectives (process evaluation);
- To what extent the chosen performance indicators were appropriate (indicator evaluation);
- To what extent the programme had an impact on development outcomes (impact evaluation).

In what follows, we will discuss evaluations of the three main types of results-based aid with a particular focus on how *impact* evaluations have (or have not) been conducted.

Output-based aid

The nature of output-based aid programmes – which involve contracts with a large number of discrete entities (e.g. firms and NGOs) for specific services (e.g. water, sanitation and telecommunications) – make these aid modalities ideal candidates for impact evaluations. It is thus surprising that so few evaluations of this type of aid have been carried out, and that IEs were not more explicitly considered and embedded in the initial strategy of the GPOBA administered by the World Bank.

Even though encouraging results were found (e.g. increased assisted deliveries, immunisation and improvements in health workers' performance) in a recent evaluation of three output-based programmes in Africa, the report was unable to establish whether the results could be attributed to the programme and how the programme compared with other types of aid (GPOBA, 2008a). Initial evaluations of GPOBA were focused on implementation challenges rather than results and impacts experienced in recipient countries.[15] A recent evaluation of the initial USAID output-based aid programme in Haiti was not able to establish impact.

In 2008, GPOBA launched two impact evaluations (GPOBA, 2008b) of a water programme and a reproductive health voucher programme in Uganda. These evaluations include the development of counterfactuals and will for the first time allow an explicit comparison of the impact of output-based aid with other aid programmes. IEs are also planned in Rwanda and the Democratic Republic of Congo. Equally, the recent RBF for Health trust fund announced that it will support learning by doing and rigorous impact evaluations.[16]

COD aid

COD aid programmes are still being developed, so result few evaluations are available.

GAVI: So far, two 'impact' evaluations have been carried out to establish the effects of the GAVI ISS programme. The first evaluation was in 2004 and included six case studies and analysis of data from 52 countries (Chee, 2004). An attempt was made to establish the impact of the ISS programme on immunisation performance. However, variability in data quality and an inability to establish sound counterfactuals made it 'impossible to attribute changes in performance of recipient countries to ISS funding'. A second evaluation carried out in 2007 tried to address this problem by developing a regression model to test the effects of ISS expenditures on immunisation rates (Chee, 2007). This

[15] See annual reports 2007 and 2008: http://www.gpoba.org/publications/annualreport.asp
[16] Report of First Meeting of the Inter-Agency Working Group on RBF (see http://go.worldbank.org/ECLIKAQ9J0).

quantitative approach was triangulated by further qualitative analysis, which included a number of control countries with similar characteristics (though not randomly assigned). This made the authors confident to conclude that 'ISS expenditure had a significant positive impact on immunization coverage'.

COD Aid for Education (pilot): The anticipation of a pilot of the COD aid programme in education has generated a number of interesting 'evaluations', as well as proposals for impact evaluations once the programme gets going. So far, discussions have included conceptual reviews and focused on what type of performance indicators should be used, and whether countries will have the capacity to generate the data needed to demonstrate progress or delivery (de Renzio, 2008; Lockheed, 2008).

Proposals to evaluate the impact of COD are also being developed. The most recent proposal (Savedoff, interview 2008) sets out to evaluate the impact of COD aid at two levels: i) to analyse the impact on donor and recipient countries' policies and actions; and ii) to evaluate the link between recipient's actions and the outcome (e.g. between changes in government policies and education outcomes).

For the evaluation of the first level of impact (from COD to action), a 'process-based' approach is suggested. This involves developing a qualitative narrative of what happened and why. The first step of the evaluation would include collection and analysis of baseline data on political economy, bureaucratic relations, governance, expenditures on education, government structure and accountability relationships, and so on. The second step would involve process tracing and the development of quasi-counterfactuals in the form of systematic assessment of how other aid modalities are operating in similar settings.

A methodology for the second level of impact (from action to outcomes) still needs to be developed and will depend on the actions that governments take as a consequence of the COD aid agreement. If the recipient responds with programmes that can be tested in a small number of schools or introduced at different times across the country, it should be possible to construct a counterfactual in the impact evaluation design and generate rigorous evidence of how and why programmes achieved what they did. If, however, government action in response to COD is at the national level and indivisible (e.g. negotiating a new deal with the teachers' union), it may be more difficult to identify appropriate counterfactuals. However, the proposal emphasises that efforts should be made to learn from other quantitative analyses of government programmes that have recently emerged.

EC results-based budget support

The majority of evaluations of EC budget support programmes have focused on *process-type questions*, such as whether it has encouraged aid effectiveness principles like country ownership (for example EEPA, 2008; Schmidt, 2006 and ECDPM, 2005) and how budget support relates to wider European Union (EU) aid policies (CIDSE, 2007). Other evaluations have considered concerns surrounding the *use of results indicators* (Oxfam, 2008). The EU results-based aid distinguishes itself from other types of aid by its reliance on performance indicators. An issue frequently highlighted in the evaluations is the lack of ownership in the indicator choice, as these are often drawn from international development objectives with little input from partner governments (Volker et al., 2005). The number of indicators, and their place in the results chain, is also examined (Adam, 2002). Finally, evaluations have analysed the process and rigour of data collection (Kanbur, 2005).

Virtually no studies have analysed the *impact* of this new form of results-based aid, hence its effectiveness in contributing to achieving development results is still unclear. However, significant progress has been made in developing a methodological framework that should allow impact evaluations of general budget support programmes, which may be of relevance to evaluation of results-based budget support.

Logframe evaluation of budget support programmes: The Evaluation Framework (EF) developed by Lawson and Booth (2004) was the first major attempt to develop a practical tool to guide budget

support evaluations at the country level, building on the work of White and Dijkstra (2003).[17] The EF was applied (with some adjustments) in a joint donor study of budget support in seven partner countries between 1994 and 2004 (IDD, 2006). This qualitative study used alternative scenarios to establish counterfactuals. However, the approach did not allow establishment of the impact of programmes on poverty reduction.

Building on the lessons from the joint evaluation and an evaluation of budget support in Ghana (ODI and GCDD, 2007), the EC recently commissioned the development of the Comprehensive Evaluation Framework (CEF) (Caputo et al., 2008). This framework presents a three step approach. The first includes the assessment of changes in government systems (public financial management and policy processes) induced by budget support. The second involves an *impact* evaluation of the outcomes and impacts of the government strategy that budget support intends to support. The third combines and compares the results of Steps 1 and 2 and explores the contribution or causal relationship between the budget support programmes and the government strategy outcomes.

In addition, the CEF explicitly addresses different options for counterfactual analysis. It spells out three different approaches to building a counterfactual that could be applied depending on circumstances: quantitative model, control groups and qualitative alternative scenarios.[18] It is noted that, because budget support interventions are not discrete or targeted at a specific group, it will be difficult to use control groups (NONIE, 2007). Despite its lower level of rigour, the CEF promotes the use of more qualitative analysis of alternative scenarios. However, it also sets out that this should not be seen as a way to avoid more rigorous analyses. Resource constraints and suitability of the methods should determine the ultimate choice of methodology.

Statistical impact evaluation: A concrete proposal for the use of more quantitative analysis of (sector) budget support is set out in Elbers et al. (2008). This proposal, which has now been tested in Dutch development programmes (IOB, 2007), modifies existing statistical techniques, which have commonly been used to evaluate more discrete project-based programmes. It proposes a 'bottom-up' approach where impact evaluations at the project level are performed in a way allowing for conclusions to be drawn at a higher aggregate level. First, a representative sample of activities by a ministry is selected. Second, impact evaluations are performed for the selected activities at the project level. Then results are aggregated to generate conclusions like 'public spending in country X reduced poverty by Y percent'. It should be noted that the result is not aid specific.

Application to results-based budget support? There is a need for further thinking and research into how the proposed IE methodologies for general budget support could be applied to the specific case of *results-based budget* support programmes. It is unclear whether, as payments become more dependent on outcomes (rather than policy measures), it will be easier or more difficult to develop the intervention logic and discern policy actions resulting from this new type of aid. The CEF methodology, for example, looks at the impact of a set of inputs, i.e. transfer of funds, policy dialogue and related conditionality, technical assistance/capacity building and budget support aligned to government policies. Some of these elements may still be present in the results-based budget support modalities (e.g. technical assistance/policy dialogue), but the approach may be more hands off and the transfer of

[17] As discussed in Nonie (2007). The methodology presents a qualitative analysis of the public expenditure process, which is assumed to be influenced by budget support through both its institutional and flow-of-funds effects. It includes a standard five-level logical sequence (inputs, immediate effects, outputs, outcomes and impacts) to establish cause and effect links and the time dimension of effects. It provides detailed guidelines for research questions and approaches at each level of the framework, based on assessing whether the expected effects of budget support are present and asking additional questions relating to attribution and the counterfactual.

[18] *Quantitative models*. This includes building a quantitative model of the macro or sectoral context in which the programme under evaluation operates, and through which it is possible to test the impact of different policies (or no policies at all), compared with those under evaluation. *Control groups*. This involves considering what happened in other areas of the same country (or other countries) with strong similarities but where different policies have been implemented. *Qualitative alternative scenarios*. This consists of building a qualitative alternative scenario as a counterfactual to the actual context under evaluation. This approach is most frequently used in budget support evaluations.

aid (e.g. variable tranche) may not be linked to a framework of conditionalities aimed at implementing ex ante the implementation of a given government programme (which defines this new type of aid). Actual flow of funds will be replaced by an incentive of future payment.

Supply and demand of IE for results-based aid

Are impact evaluations still needed? The emphasis on results-based aid has the potential to affect the demand for evaluations. Donors might come to feel that if aid is tied to outcomes there is no need for impact evaluation. However, there are at least two reasons why impact evaluations should continue to be undertaken. The first relates to attribution. Improvements in outcomes may have nothing to do with the delivery of results-based aid and impact evaluations can potentially establish this link. Second, even when success can be attributed to aid, it is still possible that the improved outcome was achieved at very high cost, and alternatives may be more cost effective. Thus, there is still an important role for impact evaluations for the purpose of learning.

An increasing demand for *impact* evaluations. The vast majority of evaluations have focused on process rather than on impact and on recording changes rather than attribution of changes to interventions. This is changing. The debate on aid effectiveness and the increased interest in results (as reflected in results-based aid) has generated a surge of interest in better evidence and formal evaluation techniques (Elbers et al., 2008).

… and at a higher level of aggregation. Aid, and results-based aid programmes in particular, is increasingly moving away from project aid towards sector and general budget support. And the impact evaluation question must be considered at a higher level of aggregation, a level for which quantitative techniques have not yet been well designed. This has led to the development of new techniques, such as 'statistical impact evaluation', as proposed by Elbers et al. (2008). This has now been tested and it can be expected that this type of evaluation will become more important in the future.

Methodology: Alternatives for RCTs. Review of the literature as well as interviews indicate growing support for approaches that try to establish impact through the development of counterfactuals. There is a concern with, on the one hand, a lack of rigour in recent qualitative analyses and, on the other, an excessive emphasis (spurred on by micro analysis) on counterfactual analysis through RCTs (Caputo, 2008). Proposals for impact evaluations of results-based government programmes (such as EC budget support and COD) involve a mixed method approach, including qualitative analysis of case histories at the policy level as well as quantitative and statistical analysis to measure programme impacts. The scope for good *quantitative* analyses of government or sector-wide programmes (as opposed to projects) has been shown to be wider than previously believed (CGD, 2006).

Moving from aid to development evaluation? Statistical impact evaluation techniques do not allow for evaluation of individual aid programmes, but rather of certain 'development' programmes or total government expenditure. Given the attempts of the donor community under the Paris Declaration to increasingly harmonise and pool aid, this approach would seem appropriate. Given this approach, however, there will be a need to increase ownership and strengthen evaluation capacity in partner countries.

Opening the black box. Quantitative models are likely to become more prominent in future evaluations of development programmes. Their main weakness is that they do not explain why interventions are effective. Proponents of statistical evaluations recognise that qualitative and descriptive studies will be necessary to complement quantitative analysis.

Data availability. One of the most significant impediments to impact evaluations has been the availability of reliable data on development outcomes. The move towards more quantitative methods of assessment will likely put greater demands on data availability. An interesting question is whether results-based aid programmes will enhance countries' capacity to generate reliable data. There is some evidence that, in countries with results-based aid programmes, data collection has significantly

improved. For example, several countries that were part of the GAVI ISS programme were found to have improved data significantly over the course of the programme. The ISS programme 'appears to have had significant impact on countries to address the problem of data quality' (Chee et al., 2007). As part of the GAVI programme, countries are required to pass a data quality audit (with score of .80) to receive reward shares. The audit is expressly geared towards improving the capacity for data collection, which can only strengthen future (impact) evaluation efforts. The choice of indicators to be monitored as part of the results-based aid has been the subject of significant debate. Available data are limited and not always best suited to properly monitor chosen outcomes (Eurodad, 2008).

Given the unique emphasis of results-based aid on countries' ability to generate their own data and monitoring systems, we can potentially expect a greater interest on the part of recipient governments in contributing to and conducting impact evaluations.

Supply driven. So far, evaluations of results-based aid have typically been carried out by external evaluators, mostly linked to donor agencies. There is a need for greater involvement of local, in-country experts in research and evaluation in the interest of capacity building and sustainability.

Conclusions and ways forward

Results-based aid is a relatively new form of aid that has generated a lot of enthusiasm among donors and development practitioners, as it provides a strong link between development interventions and outcomes. Results-based aid does not make impact evaluations redundant. There is still a need to assess whether development outcomes are generated by particular interventions, to assess the cost effectiveness of interventions and to evaluate how interventions have generated certain impacts. So far, the lack of evaluations has made it difficult to conclude whether results-based aid has created incentives for poverty reduction or had an additional impact on poverty. The tide is changing, however, and qualitative and quantitative evaluation frameworks are being developed and being tested.

The changing landscape towards more results-based approaches of aid and evaluations of these approaches has a number of implications that need to be considered:

- **Closer ex ante collaboration between operation and evaluation departments.** Results-based aid approaches have an inherent focus on results (in terms of outcomes and impacts) as disbursements are made dependent on them. This could potentially make it easier (from an outcome measurement perspective) to conduct impact evaluations. However, impact evaluations will need to be considered at the design stage of the programme. This may have implications for organisations with separate operational and evaluation departments, such as DFID. Traditionally, evaluation departments only get involved ex post and there may be a need to get involved more ex ante in the selection of indicators and development of the logical chain. A greater focus on results upstream will likely benefit impact evaluation.
- **Tools and methodology.** A number of tools have already been designed but further testing and developing of quantitative and qualitative tools and approaches will be needed. The purpose of impact evaluations will likely be less on accountability or demonstrating results[19] (as these are an inherent part of results-based aid programmes) and more on learning what does or does not work.
- **Appropriate skills.** The use of new qualitative and quantitative methodologies to evaluate the impact of aid at the programme level will require adequate skills. In recent years, two trends have taken place. First, the economics profession has been captivated by applications of RCT (applied to projects), with much more limited interest in more traditional quantitative methods, such as econometric or macroeconomic analysis (Deaton, 2008), which are now re-emerging as feasible approaches for evaluating programme-based aid. Second, the move towards programme-based approaches has led to a decline in the use of quantitative tools and skills in evaluating these types of aid. Quantitative (as well as qualitative) methods are likely to broaden to include a greater emphasis on spelling out logical results chains, statistical analysis

[19] Although attribution will still need to be established.

and cost-benefit analysis. Evaluations of results-based aid and COD aid at the programme level will require a range of qualitative and quantitative skills.

- **Statistical capacity building.** Results-based aid approaches as well as proposed 'statistical evaluation' techniques are data intensive and will require improving data collection and statistical capacity in partner governments. This may be built into results-based aid programmes. Dialogue with partner governments on the availability of evidence will need to be further strengthened.
- **Information exchange.** As new approaches and techniques are being tested, there will be a need for exchange of information on new methodologies and the effectiveness of this type of aid. This is a role that could be played by existing networks on impact evaluation such as NONIE. Interviewees emphasised the value of structured efforts to synthesise existing research and identifying best practice examples of impact evaluations.

4. Comparing sector-specific experiences with impact evaluations

A focus on sector-specific histories and dynamics of impact evaluation production, communication and use dynamics reveals a number of important similarities and differences. Similarities include a growing recognition of the need to approach impact evaluations as part of a broader monitoring and evaluation system; the importance of involving multiple stakeholders in the evaluation process to promote uptake; and the utility of exploring innovative methods to assess impact. The differences across sectors, however, appear to be starker, and are thus important in informing efforts to promote more strategic generation, communication and use of impact evaluations in the development arena. Key differences are as follows, mapped out against our hypotheses in Table 2.

History: Impact evaluations have a long history in the health and agriculture/NRM sectors, and thus a broader array of experiences from which to draw lessons, in terms of not only evidence about programme interventions that do and do not work, but also methodologies and practical implementation challenges. By contrast, experience in the social development, humanitarian and infrastructure sectors is more recent and methodological approaches are still being pioneered. In the case of results-based aid, the field is extremely new, so our analysis was more at the level of a 'think-piece'.

Suitability: Views on the suitability of impact evaluations as an evaluation approach vary considerably. In the health sector, in particular, and to a lesser degree in the agriculture/NRM sector, there is recognition that impact evaluations are strongly suited to providing robust evidence on a range of key questions in particular policy areas of the field. In the case of social development, thanks in no small part to the demonstration effect of the impact evaluations on cash transfers, there is also a growing appreciation that impact evaluations that draw on mixed methods approaches can provide valuable insights into the effectiveness of social development interventions. Views in the humanitarian and infrastructure sectors and in results-based aid are all considerably more cautious, with at best a recognition that impact evaluations (as currently conceived by donors) may provide limited purchase on key issues in the sector. For the humanitarian sector, this is particularly the case with acute emergency situations, where staff have little time to learn complex methodologies. There also appears to be a strong concern that unintended effects will be underreported or glossed over, given pressures to justify funding levels. In the case of infrastructure sector initiatives, limitations are perceived to be particularly significant in the case of large-scale infrastructure projects.

Methodological innovation: There is a strong interest and practice of methodological innovation for impact evaluations in the health and social development sectors, involving creative use and sequencing of qualitative and quantitative methods in order to unpack impact pathways. In the agriculture/NRM sector, there is also an increasing use of mixed methods approaches, but simultaneously a concern that qualitative insights are not accorded sufficient weight. Methodological innovation is more limited in the humanitarian and infrastructure sectors, although in the latter case there is considerable attention to creative approaches to ensuring variation in the unit of analysis, whether this be household, municipality or state. In results-based aid, efforts are still embryonic.

Gestation of intervention: An important challenge in the use of impact evaluations, especially in terms of learning in the context of a specific programme intervention (compared with the development of a broader global evidence base), is the length of gestation of programme interventions. In the human and social development field as well as the results-based aid sector, there is recognition that programme impacts often take several years; for agriculture/NRM and infrastructure, it may take considerably longer for impacts to be felt.

Implementing agencies: In order to promote greater stakeholder engagement and use of evaluation findings, the choice of evaluation implementation agency can be of critical importance. In the case of the health sector, evaluations can be carried out by a range of actors, including academic researchers,

international agencies and NGOs. There is also a strong trend of working with developing country governments (especially ministries of health) as well as local researchers (university or NGO based). Although impact evaluations are more fledgling in the social development sector, a similar type of trend is emerging. In the agriculture/NRM sector, a diverse array of actors tend to be involved, but the private sector plays an important role, especially in the area of seed varieties. In the case of the humanitarian, infrastructure and results-based aid sectors, evaluators tend to be international agencies.

Commissioning: The commissioning of impact evaluations tends to be supply driven in the case of all but the health and social development sectors, where there is a growing demand from developing country governments, especially in Latin America, to undertake impact evaluations. There is also a small but growing interest on the part of NGOs in the social development and humanitarian sectors to undertake impact evaluations.

Communicating findings at the national level: Communication of findings appears to be relatively robust in the health sector, facilitated by the fact that health officials are typically more accustomed to an evidence-based culture. Knowledge sharing from impact evaluations is also growing in the social development sector, especially in Latin America, India and Indonesia. But there are still important capacity gaps, especially because of the limited number of non-economist social scientists in many developing, especially low-income, countries. In the other sectors, the communication of evaluation findings seems to be limited, and in results-based aid it is still too early to assess.

Communicating findings at the international level: There is strong interest and a number of new coordinating mechanisms to communicate findings from impact evaluations in the health and social development sectors at the international level, including the Poverty Action Lab's policy brief cases and policy bulletins, the World Bank and IFPRI's impact assessment discussion papers. The Campbell and Cochrane Collaboration also provide inspiration to researchers working in the development field (Glennester, interview 2008). Similarly, there is a range of initiatives targeting academic and policymaker audiences in the CGIAR system in the agriculture/NRM and infrastructure sectors, but a number of respondents expressed concern that learning opportunities are limited by a tendency to 'smooth off critical edges' and focus predominantly on positive results.

Use of evaluation findings: In the health sector, evaluation findings are routinely used in the medical field, and there is growing uptake in the public health field, but the key challenge involves scaling up rather than innovative methods. In the social development sector, evaluation findings are starting to be taken up more readily, especially in the case of social protection interventions. Important new initiatives include the establishment of CONEVAL in Mexico, which is aiming to institutionalise systems to demonstrate learning from findings among implementing agencies and to promote their integration into subsequent workplans. There is, however, a strong awareness of the importance of promoting political will and ensuring that findings are feasible in specific political contexts. In the agriculture/NRM sector, there is a concern that use of findings is hindered by limited attention in impact evaluations to broader context and programmatic variables, especially in the case of more complex interventions. In the humanitarian sector, given the relatively small role that impact evaluations play, there is no significant concern that evaluation results are unduly biasing donor policies, whereas in the infrastructure sector there is a general perception that the use of results stops at reporting to donors and official accountability objectives, rather than being used for broader learning purposes. Finally, in the case of results-based aid there is recognition that impact evaluations will remain an important tool, but it will be critical to assess the spillover effects of this new aid modality on the monitoring and evaluation cultures of other development sectors.

Table 2: An assessment of the hypotheses on IE production and use across sectors[20]

	Social development	Humanitarian	Agriculture/NRM	Infrastructure
1. *IE (of all types) is relevant only relative to the timescale over which an intervention might plausibly affect beneficiaries, so may require long timeframes.*	2-5 years	Potentially very short term, for emergency response. Otherwise, some years (e.g. disaster recovery)	A few years for e.g. uptake of new seed variety, a few decades for e.g. NRM	E.g. 5-7 years before any observable impact can be seen as a result of a rural electrification programme
2. *Experimental IEs are most suited to interventions that have short and relatively simple impact pathways.*	Yes, but also gender empowerment, post-conflict trauma, but textbooks, worming, cash transfers	Yes, although limited number of studies so far. E.g. distributing blankets, cash	Generally yes, e.g. animal medicine. Some work on impact of research trying to go beyond simple technical 'widget' research	Randomisation tends to work well when the sample size of the unit of analysis (e.g. household) is relatively large
3. *Experimental IEs are most suitable where an intervention can be modelled as involving discrete, homogenous outputs.*	Cash, vaccines	Yes, although limited number of studies so far. E.g. distributing blankets, cash	Yes: seeds, vaccines	Yes, but some interventions (roads) can have unforeseen effects on prices and welfare outcomes – making IE difficult to conduct
4. *Experimental IEs require the intended effects of an intervention to be quantifiable.*	Yes but also combine with qualitative information on pathways and process data – strong support for this	Yes, but many outcomes not quantifiable. Multiple and mixed methods needed	Yes, e.g. productivity, nutrition	Yes, but quantitative IEs do not provide all answers. Qualitative provide answers to why and how questions
5. *Experimental IEs are only feasible in contexts where it makes sense to investigate what would have happened in the absence of the intervention.*	Yes	Yes. Often ethically inappropriate to do full RCT in emergency response interventions	Yes – not so plausible for e.g. NRM	Yes, so difficult to construct counterfactuals for mega projects like dams, bridges and subways
6. *Experimental IEs are suitable where effects are attributable to distinct forces/actions/interventions.*	Yes	Yes	Yes – so not e.g. biodiversity	Yes
7. *Where suitable, experimental IEs are able to*	Yes, a much celebrated	Potentially, in certain	On certain timescales, e.g. uptake	Quantitative IEs can

[20] The Results Based Aid case study is not included in this table as impact evaluations are only starting to emerge and there is not yet a large body of evidence available to test the hypotheses. Interventions are also often not at a project level (apart from Output Based Aid), which makes experimental approaches hard to use. For a discussion of alternative impact evaluation approaches (using counterfactual analysis), please see section 3, case 5 on Results Based Aid.

provide robust evidence proving (and quantifying) the effectiveness of a project/programme/policy against predefined goals.	case is greater school enrolment through cash transfers	contexts	of seeds.	give robust answers about whether or not a project is having the desired impact – but have a number of flaws
8. *Experimental IEs are most suited to testing the effectiveness of a small number of interventions.*	Yes: cash, textbooks, de-worming tablets	Yes, given that hypothesis 3 makes it most relevant for emergency response, which is likely to be ethically inappropriate	Yes: 'less than 25% of the sector', according to one interviewee	No, covers a wide range of interventions
9. *There are a number of potential practical issues in carrying out IEs, which (while each can be surmounted) nonetheless affects the when, where, how and by whom they are produced. Owing to these issues and the perceptions of those commissioning IEs, methodological concerns often receive disproportionate weight in deciding what and where to evaluate.*	Yes but growing methodological innovation – trauma, empowerment, anti-corruption	Too early to say	Yes, although pressure to apply to other areas e.g. impact of research, even though limited relevance here	But there are increasing pressures to apply IE methodology to suitable interventions – 'method in search of application' On other hand, findings are often seen as more important than methods
10. *Experimental IEs are more likely to be carried out when they are expected to generate positive results. Like other types of evaluation, they tend to be published only if they demonstrate positive results.*	Yes, this was rationale of cash transfers in Mexico but textbooks – negative findings	Too early to say	Limited evidence either way. By implication, possibly yes	IEs often conducted to legitimate policy decisions. IEs with poor results interrogated and generally not taken up
11. *Because IEs are still relatively new outside the health and agriculture fields, they tend to be undertaken on the basis of researcher or donor agency suggestion, rather than being demand driven.*	Yes, but important exception of Mexico, and increasingly in India and Indonesia	Too early to say, although some indications backing this up	As above	Commission of IEs tend to be driven by bilateral and multilateral donors
12. *The production of experimental IEs is largely driven by upward accountability to donors.*	To some extent, but some government demands plus also academic interest	Early indications suggest upward accountability and 'good news stories'	Yes, particularly World Bank (and through CGIAR as proxy)	JBIC conducts IEs on interventions for accountability purposes, e.g.
13. *Experimental IEs tend to be commissioned less frequently to fulfil downward accountability, or*	CONEVAL is exception that proves the rule	Yes	Yes	Yes

operational learning purposes.				
14. *There are three main channels (not mutually exclusive) as to how experimental IEs can be put to use:*				
a. *Directly: experimental IEs are a major input to managing programmes based on results. They can provide a major source of evidence to shape budget allocations among different activities, and decisions to continue/discontinue/modify/scale up a smaller (possibly pilot) project.*	Yes, very important with education (textbooks, teacher remuneration)	Too early to say. Unlikely: other factors are more important	Intention to use like this not fulfilled owing to other factors determining uptake, e.g. hype around particular technology	
b. *Legitimation: experimental IEs are used to justify the actions of an organisation, particularly in the context of fundraising efforts.*	Very important with social protection	Too early to say, but quite likely	Yes. Feeling that this occurred on a grand scale; CGIAR IEs contributing to renewed focus and funding from World Bank	Yes, more common
c. *Indirect use: experimental IEs contribute to policy and practice by building up the stock of knowledge about programmatic interventions that do or do not work in addressing particular policy and programmatic challenges. A conceptual way, creating debate and dialogue, generating increased clarity. This could be through strategic feedback, or through the knowledge generated.*	Cash transfers, gender empowerment, trauma	Too early to say. Possible	Yes in some 'good practice' examples	Unclear
15. *Of the different types of use, IEs are most frequently used for legitimation. This is largely in a 'defensive' mode, to protect funding. There is also growing indirect use.*	More diverse types of use in social development field – see above. But legitimation doesn't have to be negative – can be very important in social development sector to advance progressive social agendas in face of opposition from fiscal conservative forces	Early suggestions back this up	Confirmed	Yes
16. *Factors that affect/explain the use of IEs are:*		Too early to say		Too early to say
a. *The rigour of experimental IE may be a*	Yes and importance to			

strong force for its uptake.	get political backing			
b. *Relational factors such as trust and engagement may be a large barrier to uptake.*	Yes – various agencies trying to promote government, donor and media capacities to critically assess IEs but inadequate knowledge management is a critical barrier			
c. *The fledgling state of knowledge management may be a significant barrier to uptake.*	Yes – but significant efforts by J-PAL and Poverty Lab are starting to address this			
17. *Where policymakers view experimental IE as the 'gold standard' and funding is influenced by the production of reliable experimental IE evidence, this risks skewing policy priorities towards areas most suitable for experimental IEs.*	Gold standard provides very useful evidence if feasible in terms of political considerations and is within the scope of existing resources and capacities	Some feel they may make it harder to embed reflective practice, but possibly unlikely owing to general difficulties with using evaluations	Yes, pressure experienced, feeling that some aspects of sector and of interventions neglected	Policymakers assume IEs with poor results are bad for beneficiaries, rather than trying to improve it – and policymakers may choose to ignore, as they may think aid will be suspended IE. Conversely, policymakers will pick up IEs with positive impact

5. Conclusions and policy implications

Overall, this study has highlighted the need and growing demand for greater and more strategic coordination of impact evaluation efforts. Although there is a growing recognition of the importance of impact evaluations in assessing the effectiveness (or otherwise) of development programmes, impact evaluation's potential to shape donor investments and national-level policy decision making has yet to be realised. This owes in part to insufficient attention to diverse methodological approaches to evaluation. The findings of this report suggest strongly that in all sectors there is a strong need for critical reflection on the suitability of methods to development questions and required knowledge. As such, there is a need to invest in the development of impact evaluations informed by methodological pluralism. Other key policy implications that emerged from the study highlighted a dearth of attention and resources devoted to:

1. Strategic coordination of a broad range of policy questions and related programmes to be evaluated from across different policy sectors;
2. Funding patterns and policies and the extent to which they address weak incentive structures for researchers and implementing agencies alike;
3. Different sectoral dynamics, methodological suitability and histories;
4. Capacity strengthening of developing country researchers and end users, governments and NGOs;
5. The communication of impact evaluation findings;
6. Documentation of lessons learned in such communication and policy engagement processes, including eventual uptake.

Our conclusions and recommendations focus on five key areas that key agencies could potentially address in order to add value to the field.

1. Strategic coordination: With the exception of social protection programmes and some health and education sector interventions, impact evaluations to date have been shaped by individual donor priorities, rather than in accordance with a broader strategic framework, such as the MDGs or PRSPs. However, key informant interviews suggested that there is a need to strengthen linkages between the current focus of impact evaluations on the project level and to broader policy-level questions and challenges. Clustering of impact evaluation studies could make an important contribution here. It would also assist evaluators in assessing the value of scaling up projects, and aggregating results across evaluations. Furthermore, clustering would also decrease the costs associated with IEs through allowing coordination (rather than duplication) of data collection, and greater synergies with national data collection efforts.

Clustering initiatives would, however, need to be informed by two important contextual factors. First, it will be important to ensure that birds-eye view coordination efforts are balanced by efforts to tap community-level perspectives on priority problems. We cannot assume that researchers and international agencies alone have the relevant knowledge to shape priority-setting processes – serious efforts also need to be undertaken to solicit demand from potential programme beneficiaries as well as national and local governments. Second, impact evaluations need to be viewed as a part of broader evaluation systems and their undertaking guided by clear criteria, including innovativeness, technical and political feasibility, cost effectiveness and political will to change based on evaluation findings, as well as the possibility of going to scale.

In order to improve the geographical and thematic coverage of impact evaluations, NONIE could play a useful role in deciding on appropriate clusters of evaluations and then monitoring and evaluating their implementation. In view of existing gaps, **thematic/sectoral** clustering should be informed by an understanding of the types of questions and programmes most amenable to impact evaluation methodologies, as well as an assessment of political context and policy priorities. This would be particularly pertinent in the new area of results-based aid.

In terms of **geographical** coverage, the key concern is replicability – to what extent does an intervention work across various contexts? The IADB and World Bank have played a key role in ensuring that a significant number of impact evaluations have been carried out in Latin America. Given the relatively greater investment by donors in development interventions in South Asia and sub-Saharan Africa, there is clearly an urgent need to devote more resources to impact evaluations of programmes in these regions. DFID, for example, devotes 90% of its budget to development initiatives in low-income countries; this would suggest that a corresponding level of resources should be devoted to impact evaluations in these contexts, including overcoming the challenge of a dearth of good baseline data.

In terms of **methods**, although the literature has focused considerable attention around the strengths and weaknesses of the **'gold standard' of randomisation**, our key informant interviews emphasised the importance of moving beyond this debate and focusing greater energies on methodological pluralism. This should include a combination of quantitative and qualitative methods, multidisciplinary perspectives (beyond economists and medical epidemiologists) and analytical approaches and attention to impact, process and cost-related evidence. Cross-agency collaboration could play a key role in increasing knowledge sharing about sector-appropriate methodologies and promoting flexible but rigorous quality standard guidelines.

The study's findings suggest that stronger coordination could play an important role in **improving the overall quality of impact evaluations** as well as encouraging a broader range of actors, especially developing country governmental decision makers, service providers, researchers, NGOs and end users, to become involved. Given a considerable degree of diversity among donors in views about impact evaluations and thematic priorities, however, a **model involving low to medium levels of coordination** would likely be more feasible than joint agency evaluations. Moreover, by retaining a diversity of approaches, methodological innovation would also be more likely to thrive. Recommended coordination mechanisms include:

- Information sharing – through an electronic clearing house as well as regular engagement opportunities and an archiving system to include IE proposals and then publication of results, whether positive or negative;
- Establishment of communities of practice – sector specific, so as to promote learning at the pace of each sector and engage in-depth about appropriate methodologies and management issues (such as how to hire good evaluators, fieldwork challenges) – and across sectors, in order to promote a cross-fertilisation of ideas or what Teller (2008) terms 'trans-disciplinary policy and program evaluation research';
- Joint agency IE priority setting and, especially, closer ex ante collaboration between operation and evaluation departments in the case of results-based aid. Traditionally, evaluation departments only get involved ex post and there may be a need to get involved more ex ante in the selection of indicators and development of the logical chain;
- Coordinated policy engagement and communication efforts, including meta-analysis with cost figures, so that decision makers can balance the potential impact of the menu of options with relative costs;
- Establishing partnerships with developing country implementing agencies, and supporting centres of excellence in IE in developing countries. In the case of results-based aid in particular, this will require a strong focus on data collection and statistical analysis capacities;
- Working in partnership with national evaluation agencies. As many countries currently lack such institutes, an important first step would be to support the establishment of such institutions through country visits, peer exchange and funding support.

2. Funding: Funding policies and patterns emerged as a critical variable in shaping evaluation practice on a number of different fronts. As Jones and Young (2007) found, development research funding is an area yet to be sufficiently reoriented in line with the Paris Declaration principles of harmonisation and alignment. In order to promote greater coordination and synergies, funding policies could be shaped to improve incentive structures so as to encourage:

- The involvement of a range of different stakeholders in the evaluation process (from academic researchers to southern country governments, from NGOs to the private sector, from international agencies to communication specialists);
- The publication of both positive and negative results (funding agencies could simply mandate that all results be published in the name of learning);
- The involvement of evaluators in dissemination activities tailored towards not only academic audiences but also policy and practitioner audiences (here, sufficient time is important, but also the involvement of specialist communication skills, which researchers may lack).
- The sequencing of funding so that evaluations are better integrated into programme design and evaluation findings reflected in subsequent programme phases (here a multi-staged funding schedule could facilitate this process);
- Close engagement between evaluators and programme implementers – indeed, reconceptualising evaluators as 'embedded researchers' may help to increase closer communication and information sharing;
- Investment by Northern researchers and international agencies in capacity-building support for developing country governments, researchers and NGOs. This could also include support for graduate students to do evaluations as part of their doctoral research, which would draw in more senior academics at low cost.
- Investment in replication evaluations as well as pioneering new interventions in order to promote learning about programme interventions in diverse contexts.

3. Knowledge management: Important steps have been taken in terms of knowledge management, especially the creation of the World Bank DIME and NONIE websites. Information on past, ongoing and planned impact evaluations could be further enhanced so as to inform learning, promote transparency and better inform donor investment and national government policy decision-making processes. This should include regular updates along the lines of the basic statistics that this report has generated in terms of impact evaluations by sector/theme, geographical region, methodological approach and funding/implementing agency, as well as agreeing on a common database format. In addition, a coordinating secretariat could play a valuable role in developing overarching narratives about emerging policy messages from IEs in different thematic or sectoral areas, including regionally specific and/or cross-regional messages. This would necessitate the employment of communications professionals as well as non-academic evaluators who are willing to undertake replication studies.

Given the importance of the media in promoting broader national and international awareness about impact evaluations, a media database as well as capacity building for journalists could help to strengthen interest in evaluation efforts as well as more nuanced reporting.

Equally importantly, given the relative resource intensity of impact evaluations, there is a disappointing dearth of documentation on how findings are communicated and used. NONIE could therefore play an important role in coordinating and funding this type of documentation and analysis, drawing inspiration from the initial useful models developed in IFPRI's impact assessment discussion paper series.

4. Capacity strengthening mechanisms: A dearth of capacity to carry out and use impact evaluations among developing country policymakers and researchers is recognised as an important challenge by many key informants to tackle if the method is to become more than a Northern-driven 'tool in search of an application'. This will be particularly challenging in the case of promoting new qualitative and quantitative methodologies to evaluate the impact of aid at the programme level. In recent years, two trends have taken place. First, the economics profession has been captivated by applications of RCTs (applied to projects) with much more limited interest in more traditional quantitative methods such as econometric or macroeconomic analysis (Deaton, 2008), which are now re-emerging as feasible approaches for evaluating programme-based aid. Second, the move towards programme-based approaches has led to a decline in the use of quantitative tools and skills in evaluating these types of aid. Quantitative (as well as qualitative) methods are likely to broaden to include a greater emphasis on spelling out logical results chains, statistical analysis and cost-benefit analysis. Third, proposed

'statistical evaluation' techniques are data intensive and will require improving data collection and statistical capacity in partner governments. This may be built into results-based aid programmes. Dialogue with partner governments on the availability of evidence will need to be further strengthened.

Possible capacity development approaches include learning by doing; support for a community of practice including developing country actors; training workshops for 'educated consumers' of IE; supporting the development of national centres of excellence in IE that can partner with international agencies; peer review of proposed IE methodologies; and integrating impact evaluations into broader capacity building initiatives on evaluation methods. In the latter case, NONIE could make a valuable contribution by developing and communicating a clear framework for different types of evaluations, when they should be used, for what purpose, what stage in the project/policy cycle and their potential contribution to policy decision making. It could also address in summary form the evolution of debates around suitability and ethics so the evaluation community moves beyond currently polarised positions and develops a more nuanced common ground. This could potentially take the form of handbook of good practice targeted at commissioners of IEs and researchers. It would be important for such a venture to be informed by an understanding of different history and dynamics in different policy areas. This could facilitate the commissioning of impact evaluations by policymakers themselves and in turn the utilisation of findings.

5. Improving IE communication and uptake: Little analysis has been undertaken on how impact evaluation findings are communicated and then accessed and used by developing country end users. Key factors that appear to facilitate uptake in specific programmes based on best practice examples include making evaluation data widely available (through journals, working papers and policy briefs); the presence of high-profile issue champions of the programme in question; political will based on either interest in learning from evaluations and/or a more instrumental approach to demonstrating effectiveness; early stakeholder involvement and integration of questions about potential utilisation in the design of IE; and the dissemination of messages with clear policy implications.

There is also an agreement that evaluations are important in terms of policy transfer and building up knowledge on interventions as part of a global public resource. Here, ingredients of success include: i) a critical mass of evaluations; ii) a combination of impact and process evaluation data to unpack black box issues; iii) technical rigour; and iv) cost data.

Drawing on best practices from bridging research and policy more broadly, it will be important to improve engagement with policy and civil society end users **throughout** the evaluation process in order to improve ownership of the findings and to better meet the time pressures of the policy cycle. This should involve shaping the focus of the evaluations, discussing the preliminary conclusions so as to better understand institutional and socio-cultural dynamics and constraints that might explain the quantitative findings, as well as 'translating' academic/technical findings into non-jargonistic policy-relevant messages. There is also potential scope for employing impact evaluation approaches as a planning tool as well as an ex post tool in order to encourage programme designers and implementers to envisage the eventual impacts they hope to achieve.

References

Abelson, J. and F. Gauvin (2006) 'Assessing the Impacts of Public Participation: Concepts, Evidence and Policy Implications'. Canadian Policy Research Networks Research Report P06.

Adam, C. and Gunning, J. W. (2002) 'Redesigning the Aid Contract: Donors Use of Performance Indicators in Uganda', *World Development* vol. 30, no. 12.

Adato, M. 2008. "Combining survey and ethnographic methods to improve evaluation of conditional cash transfers" in In P. Shaffer, R. Kanbur, N. Thang and E. Bortei-Doku (eds), Special Issue *Journal of Multiple Research Approaches* (Vol. 2, No. 2).

American Evaluation Association (2003) 'Scientifically Based Evaluation Methods'. American Evaluation Association Response to US Department of Education. Notice of Proposed Priority, Federal Register RIN 1890-ZA00, 4 November 4.

Asian Development Bank (ADB) (2006) *Impact Evaluation: Methodological and Operational Issues*. Manila, Philippines: ADB.

Attanasio, O.P., C. Meghir and M. Szekely (2004) 'Using Randomised Experiments and Structural Models for Scaling up: Evidence from PROGRESA Evaluation', in F. Bourguignon and B. Pleskovic (eds) *Accelerating Development*. Annual World Bank Conference on Development Economics. Washington, DC: World Bank and OUP.

Baker, J. (2000) *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Washington, DC: World Bank.

Balthasar, A. and S. Rieder (2000) 'Learning From Evaluations'. *Evaluation* 6(3): 245-260.

Bamberger, M. (2000) 'The Evaluation of International Development Programs: A View from the Front'. *American Journal of Evaluation* 21(1): 95-102.

Bamberger, M. (2006) *Conducting Quality Impact Evaluations under Budget, Time and Data Constraints*. Washington. DC: World Bank.

Barder, O., Birdsall, N., (2006) 'Payments for Progress: A Hands-Off Approach to Foreign Aid'. CGD Working Paper Number 102, December.

Behrman, J.R. (2007) *Policy-oriented Research Impact Assessment Case Study On the International Food Policy Research Institute (IFPRI) and the Mexican Progresa Anti-Poverty And Human Resource Investment Conditional Cash Transfer Program*. Impact Assessment Discussion Paper 27. Washington, DC: IFPRI.

Benner, T., S. Mergenthaler and P. Rotman (2007) 'International Bureaucracies: The Contours of a (Re) Emerging Research Agenda'. Paper presented at a DVPW IR Section Conference.

Bennett, J., F. Lubben, S. Hogarth and B. Campbell (2005) 'Systematic Review of Research in Science Education: Rigour or Rigidity'. *International Journal of Science Education* 27(4): 387-406.

Bhola, H. (2000) 'A Discourse on Impact Evaluation: A Model and its Application to a Literacy Intervention in Ghana'. *Evaluation* 6(2): 161-178.

Bird, K. (2002) *Impact Assessment: An Overview*. London, UK: ODI.

Blalock, A.R (1999) 'Evaluation Research and the Performance Management Movement: From Estrangement to Useful Integration'. *Evaluation* 5(2): 117-149.

Bloom, H. (2006) 'The Core Analytics of Randomized Experiments for Social Research'. MDRC

Bolton P. and Ndogoni, L. (2001) Cross-Cultural Assessment of Trauma-Related Mental Illness Phase II: A Report of Research Conducted by World Vision Uganda and The Johns Hopkins University. Assessing Mental Health Impact of Transitional Populations.

Bolton, P. et al. (2007) 'Interventions for Depression Symptoms among Adolescent Survivors of War and Displacement in Northern Uganda'. *Journal of the American Medical Association* 298(5): 519-527.

Brandon, P.R. (2005) 'Using Test Standard-Setting Methods in Educational Program Evaluation: Addressing the Issue of How Good is Good Enough'. *Journal of Multidisciplinary Evaluation* 3(1): 1-29.

Bryce, J., C. Victora, J. Habicht, R. Black and R. Scherpbier (2005a). *Programmatic Pathways to Child Survival: Results of a Multi-country Evaluation of Integrated Management of Childhood Illness*. Oxford and London, UK: OUP and LSHTM.

Bryce, J, C. Victora, and MCE-ICMI Technical Advisors (2005b). *Ten Methodological Lessons from the Multi-country Evaluation of Integrated Management of Childhood Illness*. Oxford and London, UK: OUP and LSHTM.

Bryce, J. et al. (2006) 'Countdown to 2015: Tracking Intervention Coverage for Child Survival'. *The Lancet* 368(9541): 1067-1076.

Caputo, E. & A. Lawson & M. van der Linde (2008) 'Methodology for Evaluations of Budget Support Operations at Country Level'. Issue Paper for the European Commission. May 2008.

Cartwright, N. (2007) 'Are RCTs the Gold Standard'. *Biosocieties 2*(2): 11-20.

Catalytic Initiative to Save One Million Lives (CI) (2008) *Evaluating the Scale-up for Maternal and Child Survival: Putting Science to Work for Mothers and Children*. Baltimore, MD: Johns Hopkins Bloomberg School of Public Health.

Catley, A., Burns, J., Abebe, D. and Suji, O. (2008) Participatory Impact Assessment: A Guide for Practitioners. Tufts university.

Center for Global Development (CGD) (2006) 'When Will We Ever Learn? Improving Lives through Impact Evaluation'. Report of the Evaluation Gap Working Group.

Chalmers, I., M. Enkin and M. Keirse (1993) 'Preparing and Updating Systematic Reviews of Randomized Controlled Trials of Health Care'. *The Milbank Quarterly* 71(3): 411-437.

Chase, R.S (2002) 'Supporting Communities in Transition: The Impact of the Armenian Social Investment Fund'. *World Bank Economic Review* 16(2): 219-240.

Chaudhury, N. et al. (2006) 'Missing in Action: Teacher and Medical Provider Absence in Developing Countries'. *Journal of Economic Perspectives* 20(1): 91-116.

Chee, G., Hsi, N., Fields, R., Schott. W.,(2004) 'Evaluation of the First Five Years' of GAVI Immunization Services Support Funding', Abt Associates, Inc.

Chee, G., Hsi, N., Carlson, K., Chankova, S., Taylor, P.,(2007) 'Evaluation of the First Five Years' of GAVI Immunization Services Support Funding', Abt Associates, Inc.

CIDSE - Cooperation Internationale pour le Developpment et la Solidarite's (2007) 'The EU's Footprint in the south', March Clegg, M. (2002) 'Commissioning Evaluation: Is There An Enlightened Approach?' Proceedings of the UK Evaluation Society Conference, December.

Cohen, J. (2008) 'Control Freaks. Are 'Randomised Evaluations' a Better Way of Doing Aid and Development Policy'. *The Economist*, 12 June.

Court, J., I. Hovland and J. Young (eds) (2005) *Bridging Research and Policy in Development: Evidence and the Change Process*. London, UK: ITDG Publishing.

Davies, P. (2004) 'Is Evidence-based Government Possible?' Paper presented at the Jerry Lee Lecture 2004, Campbell Collaboration Colloquium, Washington, DC, 19 February.

de Kemp, A. (2008) *Analysing the Effectiveness of Sector Support: Primary Education in Uganda and Zambia*. Working Paper 5. Washington, DC: NONIE.

De Renzio, P. and Woods, N. (2008) 'The Troble with Cash on Delivery Aid: A Note on its Potential Effects on Recipient Country Institutions'. Note prepared for CGD initiative on cash on delivery.

Deaton, A. (2009) Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development, NBER Working Paper No. 14690

Development Assistance Committee (DAC) (2001) *Evaluation Feedback for Effective Learning and Accountability*. Paris, France: OECD-DAC.

Dixon-Woods, M., S. Bonas, A. Booth, D. Jones, T. Miller, A. Sutton, R. Shaw, J. Smith and B. Young (2006) 'How Can Systematic Reviews Incorporate Qualitative Research? A Critical Perspective'. *Qualitative Research* 6(1): 27-44.

Duflo, E. (2004) 'Scaling up and Evaluation', in F. Bourguignon and B. Pleskovic (eds) *Accelerating Development*. Annual World Bank Conference on Development Economics. Washington, DC: World Bank and OUP.

Duflo, E. and M. Kremer (2003) 'Use of Randomization in the Evaluation of Development Effectiveness'. Paper presented at the World Bank OED Conference on Evaluation and Development Effectiveness, Washington, DC, 15-16 July.

Duflo, E, R. Hanna and S. Ryan (2008) *Monitoring Works: Getting Teachers to Come to School*. Discussion Paper 6682. Cambridge, MA: J-PAL.

Dugger, C. (2004) 'World Bank Challenged: Are Poor Really Helped?' *The New York Times*, 28 July.

Earl, S., Carden, F. and Smutylo, T. (2001) *Outcome Mapping: Building learning and reflection into development programs*. Ottowa: IDRC

Elbers, C., Gunning, J., & Hoop, K., (2008) 'Assessing Sector-wide Programs with Statistical Impact Evaluation: A Methodological Proposal' World Development, Vol. 20, No. 10.

Eurodad (2008) 'Outcome Based Conditionality: Too Good to be True'. Eurodad Report. Brussells: Eurodad.

Europe External Policy Advisor's (2008) 'Administering Aid differently: A review of the European Commission's General Budget Support', March.

European Commission (2005) 'EC Budget Support: An Innovative Approach to Conditionality', February.

European Centre for Development Policy Management (2005) 'EC Budget Support, thumbs up or thumbs down', Discussion Paper, March.

European Evaluation Society (EES) (2007) *The Importance of a Methodologically Diverse Approach to Impact Evaluation – Specifically with Respect to Development Aid and Development Anterventions*. EES Statement. Nijkerk, Netherlands: EES secretariat.

Ezemenari, K., A. Rudqvist and K. Subbarao (1999) 'Impact Evaluation: A Note on Concepts and Methods'. World Bank Poverty Reduction and Economic Management Network.

Fiszbein, A. (2006) 'Development Impact Evaluation: New Trends and Challenges'. *Evidence and Policy* 2(3): 385-393.

Foresti, M. (2007) 'A Comparative Study of Evaluation Policies and Practices in Development Agencies'. Report for AFD Evaluation Department.

Forss, K. and S. Bandstein (2008) 'Evidence-based Evaluation of Development Cooperation: Possible? Feasible? Desirable?' *IDS Bulletin* 39(1): 82-89.

Gough, D.A. (2004) 'Systematic Research Synthesis to Inform the Development of Policy and Practice in Education', in G. Thomas and R. Pring (eds) *Evidence-based Practice*. Buckingham, UK: Open University Press.

GPOBA - Global Partnership on Output Based Aid (2008a). 'Performance Based Contracting in Health: The Experience of three Project in Africa.' OBA approaches No. 19.

GPOBA - Global Partnership on Output Based Aid (2008b). 'Annual Report'.

Greenhalgh, T., G. Robert, P. Bate, O. Kyriakidou, F. McFarlane and R. Peacock (2004) 'How to Spread New Ideas: A Systematic Review of the Literature on Diffusion, Dissemination and Sustainability of Innovations in Health Service Delivery and Organisation'. Report for the NCCSDO.

Grossman, J. and F.J. Machenzie (2005) 'The Randomised Controlled Trial: Gold Standard or Merely Standard?' *Perspectives in Biology and Medicine* 48(4): 516-534.

Gunning, J.W. (2006) 'Aid Evaluation: Pursuing Development as if Evidence Matters'. *Swedish Economic Policy Review* 13: 145-163.

Habicht, J.P., C.G. Victora and J.P. Vaughan (1999) 'Evaluation Designs for Adequacy, Plausibility and Probability of Public Health Programme Performance and Impact'. *International Journal of Epidemiology* 28(1): 10-18.

Hirschmann, D. (2002) 'Thermometer Or Sauna? Performance Measurement and Democratic Assistance in USAID'. *Public Administration* 80(2): 235-255.

Hoekstra, E.J. et al. (2006) 'Reducing Measles Mortality, Reducing Child Mortality'. *The Lancet* 368(9541): 1050-1052.

Holzmann, R. (ed.) (2008) 'Social Protection and Labour at the World Bank, 2000-2008: An Overview', in World Bank (2008) *Social Protection and Labour at the World Bank, 2000-2008*. Washington, DC: World Bank.

House, E. (2003) 'Bush's Neo-Fundamentalism and the New Politics of Evaluation'. Studies in Educational Policy and Educational Philosophy 2. www.upi.artisan.s.

Independent Evaluation Group of the World Bank (IEG) (2006) Impact Evaluation: An overview and some issues for discussion. OECD DAC

Institute for International Programs (2008) 'Evaluating the Scale-up for Maternal and Child Survival: Putting Science to Work for Mothers and Children'. Prepared by IIP, Johns Hopkins Bloomberg School of Public Health for CI.

Institute of Medicine (2007a) 'Design Considerations for Evaluating the Impact of PEPFAR: Workshop Summary'. http://books.nap.edu/catalog.php?record_id=12147#toc.

Institute of Medicine (2007b) 'PEPFAR Implementation: Progress and Promise. Committee for the Evaluation of the President's Emergency Plan for AIDS Relief (PEPFAR) Implementation (2007)'. http://www.nap.edu/catalog.php?record_id=11905#toc.

Institute of Medicine (2007c) 'Plan for a Short-Term Evaluation of PEPFAR Implementation: Letter Report No. 1'. http://books.nap.edu/openbook.php?record_id=11472&page=R1.

International Development Department (IDD) and Associates (2006) 'Evaluation and General Budget Support: Synthesis Report'. Birmingham: IDD and Associates.

IOB (2007). 'Water Supply and Sanitation Programmes in Shinyanga Region, Tanzania 1990-2006.' Policy and Operations Evaluation Department, No 305.

Ito, S. N. Kobayashi and Y. Wada (2008) 'Learning to Evaluate the Impact of Aid'. *IDS Bulletin* 39(1): 71-81.

Jack, A. (2008) 'Incentives Nudge Mexico's Poor in Right Direction'. *The Financial Times*, 10 August.

Johnston, D. (2006) 'Lessons to be Learnt? The Role of Evaluations of Active Labour Market Programmes in Evidence-based Policy Making'. *Public Administration and Development* 26(4): 329-339.

Jones, H. (2006) 'A Defence of Free Will'. MPhil thesis. Mimeo.

Jones, N. and Young, J. (2007) "Setting The Scene: Situating DFID's Research Funding Policy and Practice in an International Comparative Perspective". A scoping study commissioned by DFID Central Research Department.

Jones, N., H. Jones and C. Walsh (2008) *Political Science? Strengthening Science–Policy Dialogue in Developing Countries*. Working Paper 294. London, UK: ODI.

Jones, H. (2009) 'The 'gold standard' is not a silver bullet', ODI Opinion piece 127. London: ODI

Jones, H., K. Higgins and K. Bird (forthcoming) *Equity in Development: Why It Is Important and How to Achieve It*. Working Paper. London, UK: ODI.

Kanbur, R., (2005) 'Reforming the formula, a modest proposal for introducing Development Outcomes in IDA Allocation Procedure', January.

Kelley, T., J. Ryan and H. Gregersen (2008) 'Enhancing Ex Post Impact Assessment of Agricultural Research: The CGIAR Experience'. *Research Evaluation* 17(3): 187-199.

Kremer, M. (2008) 'The Wisest Investment We Can Make: Using Schools to Fight Neglected Diseases'. Global Health Policy blog, 20 February. http://blogs.cgdev.org/globalhealth/2008/02/the_wisest_investmen_1.php.

Kremer, M., E. Miguel and R. Thornton (2005) 'Incentives to Learn'. *Education Next* 4(2): 1-9.

Krueger, A. (2002) 'Putting Development Dollars to Use, South of the Border'. *The New York Times*, 2 May.

Kuruvilla, S., N. Mays and G. Walt (2007) 'Describing the Impact of Health Services and Policy Research'. *Journal of Health Services Research Policy* 12(1): 23-31.

Lawson, A., & Booth, D., (2004) 'Evaluation Framework for General Budget Support'. Report to management group for the joint evaluation of general budget support, Overseas Development Institute, May 2004.

Leach, M. and I. Scoones (2006) *The Slow Race: Making Technology Work for the Poor*. London, UK: DEMOS.

Lee, N. and C. Kirkpatrick (2006) 'Integrated Impact Assessment: Evidence-based Policy-making in Europe: An Evaluation of European Commission Integrated Impact Assessments'. *Impact Assessment and Project Appraisal* 24(1): 23-33.

Levine, D (2006) *Learning What Works – and What Doesn't: Building Learning into the Global Aid Industry*. Working Paper. Washington, DC: CGD.

Levine, R. and W. Savedoff (2006) 'The Evaluation Agenda', in N. Birdsall (ed.) *Rescuing the World Bank: A Working Group Report and Selected Essays*. Washington, DC: CGD.

Lockheed, M., (2008) 'Measuring Progress with Tests of Learning: Pros and Cons for Progress-Based Aid in Education', CGD Working Paper No. 147, June.

Lu, C. et al. (2006). 'Effect of the Global Alliance for Vaccines and Immunisation on Diphtheria, Tetanus and Pertussis Vaccine Coverage: An Independent Assessment'. *The Lancet* 368(9541): 1088-1095.

Mackay, R. and D. Horton (2002) 'Expanding the Use of Impact Assessment and Other Types of Evaluation'. Paper presented at the International Conference on Impact of Agricultural Research and Development, San Jose Costa Rica, 4-7 February.

Mackay, R. and D. Horton (2003) 'Expanding the Use of Impact Assessment and Evaluation in Agricultural Research and Development'. *Agricultural Systems* 78(2): 143-165.

Maredia, M., D. Byerlee and J. Anderson (1996) 'Ex Post Evaluation of Economic Impacts of Agricultural Research Programs: A Tour of Good Practice'. Paper presented to Workshop on the Future of Impact Assessment in CGIAR: Needs, Constraints, and Options, Rome, 3-5 May.

Marra, M. (2000) 'How Much Does Evaluation Matter: Some Examples of the Utilisation of the Evaluation of the World Bank's Anti Corruption Activities'. *Evaluation* 6(1): 22-36.

Marra, M. (2004) 'The Contribution of Evaluation to Socialisation and Externalisation of Tacit Knowledge: The Case of the World Bank'. *Evaluation* 10(3): 263-283.

Mather, D.B., S.D. Sullivan, D. Augenstein, D.S.P. Fullerton and D. Atherly (1999) 'Incorporating Clinical Outcomes and Economic Consequences into Drug Formulary Decisions: A Practical Approach'. *American Journal of Managed Care* 5(3): 277-285.

McDavid, J. (1998) 'Linking Program Evaluation and Performance Measurement: Are There Ways We Can Build and Sustain Performance Measurement Systems?' Speaker's notes for a presentation given to the Performance Measurement Resource Team, Victoria BC, 6 May.

McDavid, J. (2005) 'Applying Qualitative Evaluation Methods', in J. Mcdavid et al. *Program Evaluation and Performance Measurement*. Thousand Oaks, CA: Sage Publications.

Michaelowa, K. and A. Borrmann (2005) *What Determines Evaluation Outcomes? Evidence from Bi- and Multilateral Development Cooperation*'. Hamburg, Germany: HWWA.

Miguel, E. and M. Kremer (2002) 'Worms: Education and Health Externalities in Kenya'. May.

Morrison, A., M. Ellsberg and S. Bott (2007) 'Addressing Gender-Based Violence: A Critical Review of Interventions'. *World Bank Observer* 22(1): 25-51.

Morrison, K. (2001) 'Randomised Controlled Trials for Evidence-based Education: Some Problems in Judging "What Works"'. *Evaluation and Research in Education* 15(2): 69-83.

Muralidharan, K and K. Sundararaman (2008) 'Teacher Performance Pay: Experimental Evidence from India'. http://www.columbia.edu/~ws2162/seminar/Muralidharan.pdf.

Naudet, J.D. and J. Delarue. (2008) *Fostering Impact Evaluations at Agence Française de Développement: A Process of In-house Appropriation and Capacity-Building*. Working Paper 2. Washington, DC: NONIE.

Newburn, T. (2001) 'What Do We Mean by Evaluation'. *Children and Society* 15(1): 5-13.

Newcomer, K. (2008) *Achieving Real Improvement in Federal Policy and Program Outcomes: The Next Frontier*. Washington, DC: George Washington University and NAPA.

Newman et al. (2002) 'An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund'. *World Bank Economic Review* 16(2): 241-274.

NONIE (2007). Impact Evaluation of New Aid Instrument and Country Programs. Discussion Paper.

Oakley, A. (1998) 'Experimentation and Social Interventions: A Forgotten but Important History'. *British Medical Journal* 317: 1239-1242.

Oakley, A. (2000) 'A Historical Perspective on the Use of Randomised Trials in Social Science Settings'. *Crime and Delinquency* 46(3): 315-329.

Oakley, A. (2006) 'Resistances to "New" Technologies of Evaluation: Education Research in the UK as a Case Study'. *Evidence and Policy* 2(1): 63-87.

Oakley, A., V. Strange, T. Toroyan, M. Wiggins, I. Roberts and J. Stephenson (2003) 'Using Random Allocation to Evaluate Social Interventions: Three Recent UK Examples'. *Annals of the American Academy of Political and Social Science* 589(1): 170-189.

ODI/CDD (2007), 'Joint Evaluation of Multi-Donor Budget Support to Ghana, Based on OECD-DAC Methodology'.

Orr, L. (1999) 'Why Experiment? The Rationale and History of Social Experiments', in *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage Publications.

Oxfam (2008) Fast Forward, How the European Commission can take the lead in providing high quality budget support' Briefing Paper, May.

Palumbo, D.J. and P.J. Wright (1980) 'Decision Making and Evaluation Research'. *Policy Studies Journal* 8(7): 1170-1177.

Patton, M. (1975) 'In Search of Impact: An Analysis of the Utilisation of Federal Health Evaluation Research'. Centre for Social Research, Minnesota University, Minneapolis.

Paul, E. (2005) 'Evaluating Fair Trade as a Development Project: Methodological Considerations'. *Development in Practice* 15(2): 134-150.

Pawson, R. (2002) 'Evidence-based Policy: The Promise of "Realist Synthesis"'. *Evaluation* 8(3): 340-358.

Paxson, C. and N.R. Schady (2002) 'The Allocation and Impact of Social Funds: Spending on School Infrastructure in Peru'. *World Bank Economic Review* 16(2): 297-319.

Petrosino, A. (2003) 'Estimates of Randomized Controlled Trials across Six Areas of Childhood Intervention: A Bibliometric Analysis'. *Annals of the American Academy of Political and Social Science* 589(1): 190-202.

Picciotto, R. (2005) 'The Value of Evaluation Standards: A Comparative Assessment'. *Journal of MultiDisciplinary Evaluation* 3: 30-59.

Pomares, J. and N. Jones (Forthcoming). 'Evidence-based policy: Similarities and Differences Across Policy Sectors. ODI Working Paper: London, UK'.

Poverty Action Lab (2008) 'Solving Absenteeism, Raising Test Scores'. Policy Briefcase 6. http://www.povertyactionlab.org/papers/briefcase06.pdf.

Pradhan, M. and L.B. Rawlings (2002) 'The Impact and Targeting of Social Infrastructure Investment: Lessons from the Nicaraguan Social Fund'. *World Bank Economic Review* 16(2): 275-295.

Pratham (2007) 'Annual Status of Education Report'. http://www.prathamusa.org/dnn/ASER2007/tabid/99/Default.aspx.

Pritchett, L. (2002) 'It Pays To Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation'. *Journal of Policy Reform* 5(4): 251-269.

Proudlock, K. and Ramalingam, B. (2008) 're-thinking the impact of humanitarian aid: background paper for the 24th ALNAP Biannul. London: ALNAP

Prowse, M. (2007) 'Aid Effectiveness: The Role of Qualitative Research in Impact Evaluation'. ODI Background Note, December.

Prowse, M. (2008) 'Impact Evaluations and Interventions to address Climate Change'. Scoping Study commissioned by the International Initiative for Impact Evaluation.

Raitzer, D. and K. Winkel (2005) *Donor Demands and Uses for Evidence of Research Impact: The Case of* CGIAR. Washington, DC: CGIAR Science Council.

Ramalingam, B. and H. Jones with J. Young and T. Reba (2008) *Exploring the Science of Complexity: Ideas and Implications for Development and Humanitarian Efforts*. Working Paper 285. London, UK: ODI.

Ravallion, M. (2005) 'Evaluating Anti-Poverty Programs'. Policy Research Working Paper 3625. Washington, DC: World Bank.

Ravallion, M. (2008) *Evaluation in the Practice of Development*. Policy Research Working Paper 4547. Washington, DC: World Bank.

Riddell, R. (2008) 'Measuring Impact: The Global and Irish Aid Programme'. Final report to the Irish Aid Advisory Board.

Roche, C. (2000) 'Impact Assessment: Seeing the Wood and the Trees'. *Development in Practice* 10(3&4): 543-555.

Ruprah, J. (2008) 'You Can Get It If You Really Want: Impact Evaluation Experience of the Office of the Evaluation and Oversight of the Inter-American Development Bank'. *IDS Bulletin* 39(1): 23-35.

Ryan, J. and X. Meng (2004) *The Contribution of IFPRI Research and the Impact of the Food for Education Program in Bangladesh on Schooling Outcomes and Earnings*. Discussion Paper. Washington, DC: IFPRI.

Sanderson, I. (2002) 'Evaluation, Policy Learning and Evidence-based Policy Learning'. *Public Administration* 80(1): 1-22.

Sandison, P. (2005) 'The Utilisation of Evaluations', in ALNAP *Review of Humanitarian Action in 2005*. London, UK: ALNAP.

Save the Children (2009) 'Effective programmes; lesson learning and impact', taken from website on 13th March 2009 http://www.savethechildren.org.uk/en/31_56.htm#LEARNING

Schaul, M. (2001) 'Early Childhood Programs: The Use of Impact Evaluations to Assess Program Effects'. Report to the Chairman, Subcommittee on Oversight of Government Management, Restructuring and the District of Columbia: Committee of Governmental Affairs, US Senate.

Schmidt, P., (2006), 'Budget support in the European Commission's Development Cooperation'.

Schorr, L. (2003) *Determining 'What Works' in Social Programs and Social Policies: Toward a More Inclusive Knowledge Base.* Washington, DC: The Brookings Institution.

Science Council (2006a) 'CIPs contribution to the Genetic Improvement of Potato'. CGIAR Science Council Brief 5. Standing Panel on Impact Assessment.

Science Council (2006b) 'Costs and Benefits of CGIAR-NARS Research in Sub-Saharan Africa'. CGIAR Science Council Brief 9. Standing Panel on Impact Assessment.

Science Council (2006c) 'Impact Assessment of Policy-Oriented Research in the CGIAR: A Scoping Study Report'. CGIAR Science Council.

Science Council (2006d) 'Spillover Increases Returns to Sorghum Genetic Enhancement'. CGIAR Science Council Brief 4. Standing Panel on Impact Assessment.

Science Council (2006e) 'Policy-Oriented Research in the CGIAR'. CGIAR Science Council Brief 18. Standing Panel on Impact Assessment.

Science Council (2006f) 'Improved Tilapia Benefits Asia'. CGIAR Science Council Brief 6. Standing Panel on Impact Assessment.

Science Council (2006g) 'The Impact of Modern Rice Varieties on Livelihoods in Bangladesh'. CGIAR Science Council Brief 8. Standing Panel on Impact Assessment.

Science Council (2008h) 'Impact of Agricultural Research in South Asia since the Green Revolution'. CGIAR Science Council Brief 21. Standing Panel on Impact Assessment.

Science Council (2006) 'Impacts of a 'Food for Education' Program in Bangladesh'. CGIAR Science Council Brief 3. Standing Council on Impact Assessment.

Science Council (2006) 'Impacts of International Wheat Breeding in the Developing World'. CGIAR Science Council Brief 7. Standing Panel on Impact Assessment.

Sherman, L.W. and H. Strang (2004a) 'Experimental Ethnography: The Marriage of Qualitative and Quantitative Research'. *Annals of the American Academy of Political and Social Science* 595(1): 170-189.

Sherman, L.W. and H. Strang, H.(2004b) 'Verdicts or Inventions? Interpreting Results from Randomized Controlled Experiments in Criminology'. *American Behavioral Scientist* 47(5): 575-607.

Smutylo, T. (2001) 'Crouching impact, hidden attribution: Overcoming Threats to Learning in Development Programs', Draft Learning Methodology Paper, IDRC Evaluation unit.

Stiglitz, J. (1998) 'Towards a New Paradigm for Development: Strategies, Policies, and Processes'. Prebisch Lecture, 19 October.

Stoker, G. and S. Greasley (2005) 'The Case for an Experimental Approach in Applied Social Research: An Illustration from the Area of Civil Renewal Policy'. http://www.ipeg.org.uk/papers/case_for_experiments.pdf.

Teller, C.H. (2008) 'Lost Opportunities and Constraints in Producing Rigorous Evaluations of USAID Health Projects, 2004-2007'. *IDS Bulletin* 39(1): 90-97.

The Financial Times (2008) 'Cash for Safe Sex'. *The Financial Times*, 25 April.

Tudur, C., P.R. Williamson, S. Khan and L.Y. Best (2000) 'The Value of Aggregate Data Approach in Meta-analysis with Time-to-event Outcomes'. Journal of the Royal Statistical Society 164(2): 357-370.

van de Putte, B. (2001) 'Follow-up to Evaluations of Humanitarian Programmes'. Findings of the ALNAP Commissioned Study: Improving Follow-up to Evaluations of Humanitarian Programmes. Paper submitted to the ALNAP Biannual Meeting, 26-27 April.

Victora, C., J. Habicht and J. Bryce (2004) 'Evidence-based Public Health: Moving beyond Randomised Trials'. *American Journal of Public Health* 94(3): 400-405.

Volker, H., Hasse, O. and Koppensteiner, M. (2005) 'EC Budget Support: Thumbs Up or Down?'. *ECDPM Discussion Paper 63.* Maastricht: ECDPM.

Watts, J. D. Horton, B. Douthwaite, R. La Rovere, G. Thiele, S. Prasad and C. Staver (2007) 'Transforming Impact Assessment: Beginning the Quiet Revolution of Institutional Learning and Change'. *Experimental Agriculture* 44(1): 21-35.

Weiss, C. H. (1999) 'The Interface between Evaluation and Public Policy'. *Evaluation* 5(4): 468-486.

White, H. (2005) *Challenges in Evaluating Development Effectiveness*. Brighton, UK: IDS.

White, H. (2008) *Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation*. Working Paper 7. Washington, DC: NONIE.

White, H. and M. Bamberger (2008) 'Introduction: Impact Evaluation in Official Development Agencies'. *IDS Bulletin* 39(1): 1-11.

Wiebe, F.S. (2008) *Aid Effectiveness: Putting Results at the Forefront: MCC's New Institutional Approach*. Working Paper. Washington, DC: MCC.

Will, C.M. (2007) 'The Alchemy of Clinical Trials'. *BioSocieties* 2(1): 85-99.

World Bank (2003) 'Project Performance Assessment Report Bangladesh. Female Secondary School Assistance Project'. http://www.wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2003/07/28/000160016_20030728100253/Rendered/PDF/262260BD0PPAR.pdf.

World Bank (2007) *World Development Report 2008: Agriculture for Development*. Washington, DC: World Bank.

World Bank (2005) *Tools for Institutional, Political and Social Analysis (TIPS): A Sourcebook Poverty and Social Impact Analysis (PSIA), Volume 1*. Washington, DC: World Bank.

World Vision (2005) 'Learning through Evaluation with Accountability and Planning (LEAP): World Vision's Approach to Design, Monitoring and Evaluation'. http://www.worldvision.org.uk/upload/pdf/LEAP_Summary_Edition.pdf.

Yau, P. (2007) 'Working Toward a Better Environment'. *China Daily*, 22 November.

Appendix 1: Impact evaluation database overview and findings

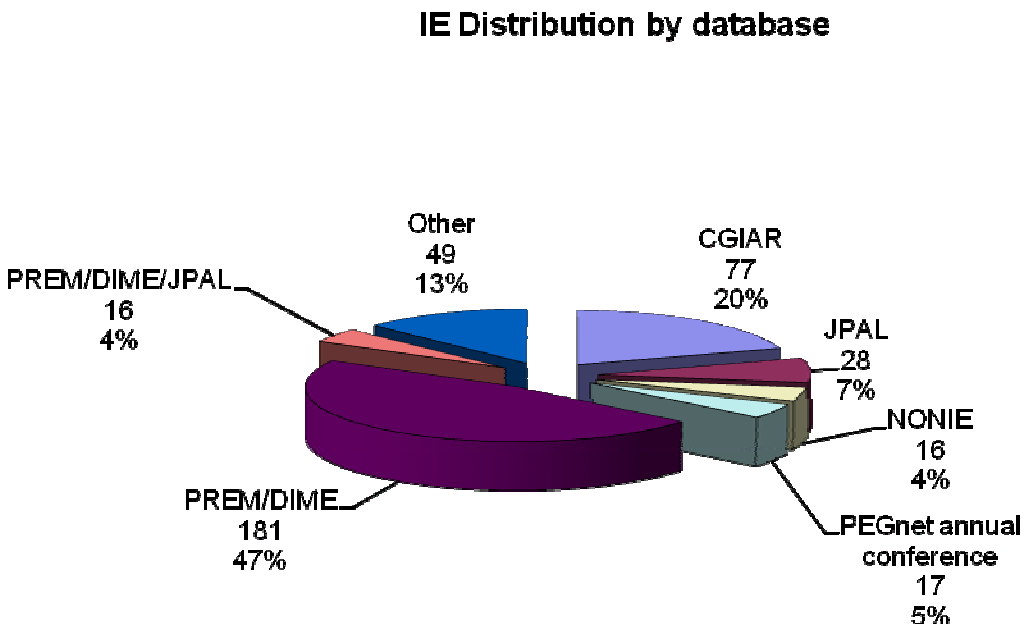
Overview

The database constructed during the first scoping study (drawing on publicly available IEs from 4 databases – World Bank DIME, NONIE, PREM and J-PAL) was enhanced by including studies from the CGIAR database of impact assessments, those presented at a PEGnet annual conference held in Accra, Ghana in 2008 and conducting a hand search for IEs in the following sectors: governance; post conflict; public sector management (e.g. decentralisation, political reservations); rural development; urban development; infrastructure (e.g. roads); private sector development; health (e.g. de-worming, HIV and AIDS, healthcare services, subsidisation of public health goods). Only those studies conforming to our definition of an IE (a study assessing a counterfactual – implicitly or explicitly – with a focus on final welfare outcomes, using qualitative, quantitative or mixed methods) were included. Any duplicate IE studies (for example from different databases) were deleted.

Database coverage

As a result, the number of IEs in the annotated database increased from 250 to 350. Chart 1 shows which databases the IEs in our annotated database come from. The largest proportion of studies comes from the PREM/DIME database (47%). CGIAR follows with 20%, studies found across other websites (such as the USAID clearing house) represent 13% of our database, J-PAL represents 7% of the IEs while PREM/DIME/J-PAL and NONIE, both contribute 4% to the database.

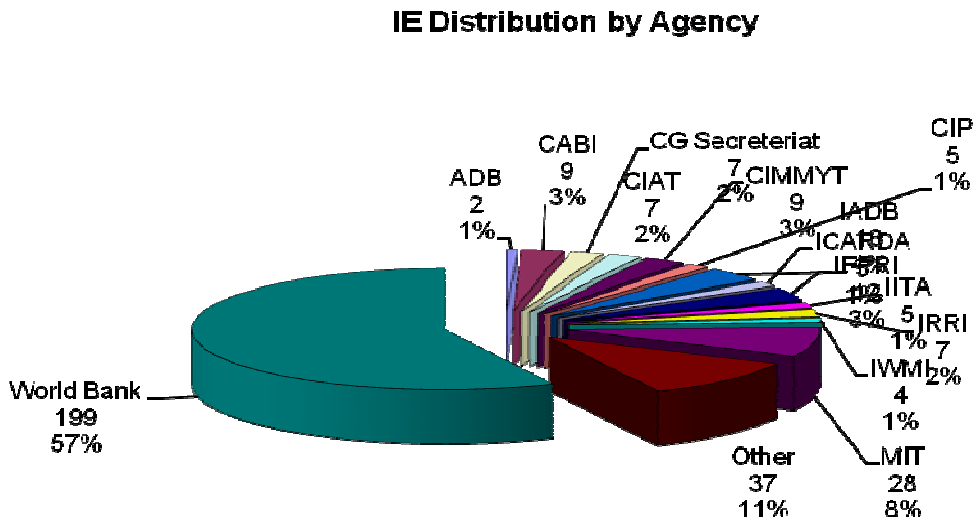
Chart 1: IE distribution by database



Agency coverage

Chart 2 below shows the World Bank is the undisputed leader in the field of IEs, followed by MIT and the Inter-American Development Bank. CGIAR research centres, such as CIAT, CIMMYT, CIP, ICARDA, IFPRI and others in total make up a significant proportion of studies. Although studies found across the PREM/DIME/J-PAL websites are classified as World Bank-implemented IEs (i.e. WB as the agency), the World Bank's role in relation to these is rather heterogeneous. The Bank's involvement has ranged from direct undertaking of the impact evaluation intervention to providing support in its implementation, involving a variety of other academic and development institutions.

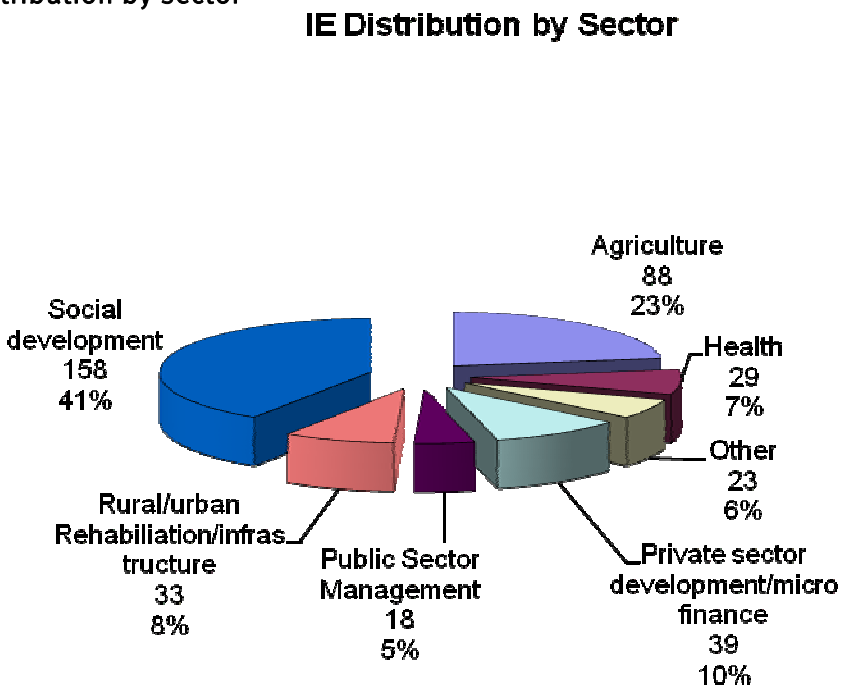
Chart 2: IE distribution by agency



Sectoral coverage

Chart 3 shows the largest proportion of impact evaluations in our database have been carried out in the Social Development sector (41%). This is followed by Agriculture (23%), Private Sector Development/Microfinance (10%), Urban-Rural Development and Infrastructure (8%), Health (7%), other (7%), and Public Sector Management (5%).

Chart 3: IE distribution by sector



For each sector, chart 4 offers a further breakdown of the total number of IE studies across more specific areas of interest, providing a sense of the most highly concentrated subcategories, as follows:

- Social Development: Social Protection and Education

- Private Sector Development/Microfinance: Microfinance and Privatisation (of public water provision mainly)
- Urban-Rural Development and Infrastructure: Infrastructure i.e. Roads/Irrigation
- Public Sector Management: Decentralisation and Political Reservations
- Health: Deworming, HIV/AIDS, Healthcare services, Subsidization of Public Health Goods
- Agriculture: research and technologies

Overall and within Social Development, **Social Protection** and **Education** are the subcategories with the highest percentage of studies. Studies within Agricultural research rank next, followed by those in agricultural technologies and microfinance. Studies in infrastructure and urban development come next. Chart 6 illustrates how, within the Social Protection subcategory, the most common evaluations are those focusing on the impact of Cash Transfers (46%), and the effects on a variety of outcomes ranging from consumption, health, nutrition and schooling to poverty and inequality. Furthermore, a single programme, the Mexican conditional cash transfer program Progresa, accounts for approximately 20% of the total number of IEs in Social Protection. IEs related to Labour Markets and Employment (21%), and Social Funds (16%) follow. The former generally consists of public work programs introduced by governments aiming at responding to rising levels of unemployment (e.g. Active Labor Market Programmes in Eastern Europe). The latter include poverty-alleviating funds that generally aim at enhancing the population's access to basic services (e.g. Bolivian Investment Fund, Emergency Social Investment Fund of Nicaragua, Honduras Social Investment Fund). Social Insurance accounts for 8% of the studies in social protection and draws extensively on the Chinese and Vietnamese government-implemented health insurance schemes. IEs of Childcare programmes, Food Aid and Food Transfers together add up to 9% of the total number of studies. [21]

[21] Our study classifies as 'Childcare programmes' all governmental interventions supporting child development which are not cash transfers (e.g.The 1991 Hogares Comunitarios Program (HCP) in Guatemala which provided affordable and reliable childcare alternatives to working parents).

Chart 4: No. of evaluations by sub-sector

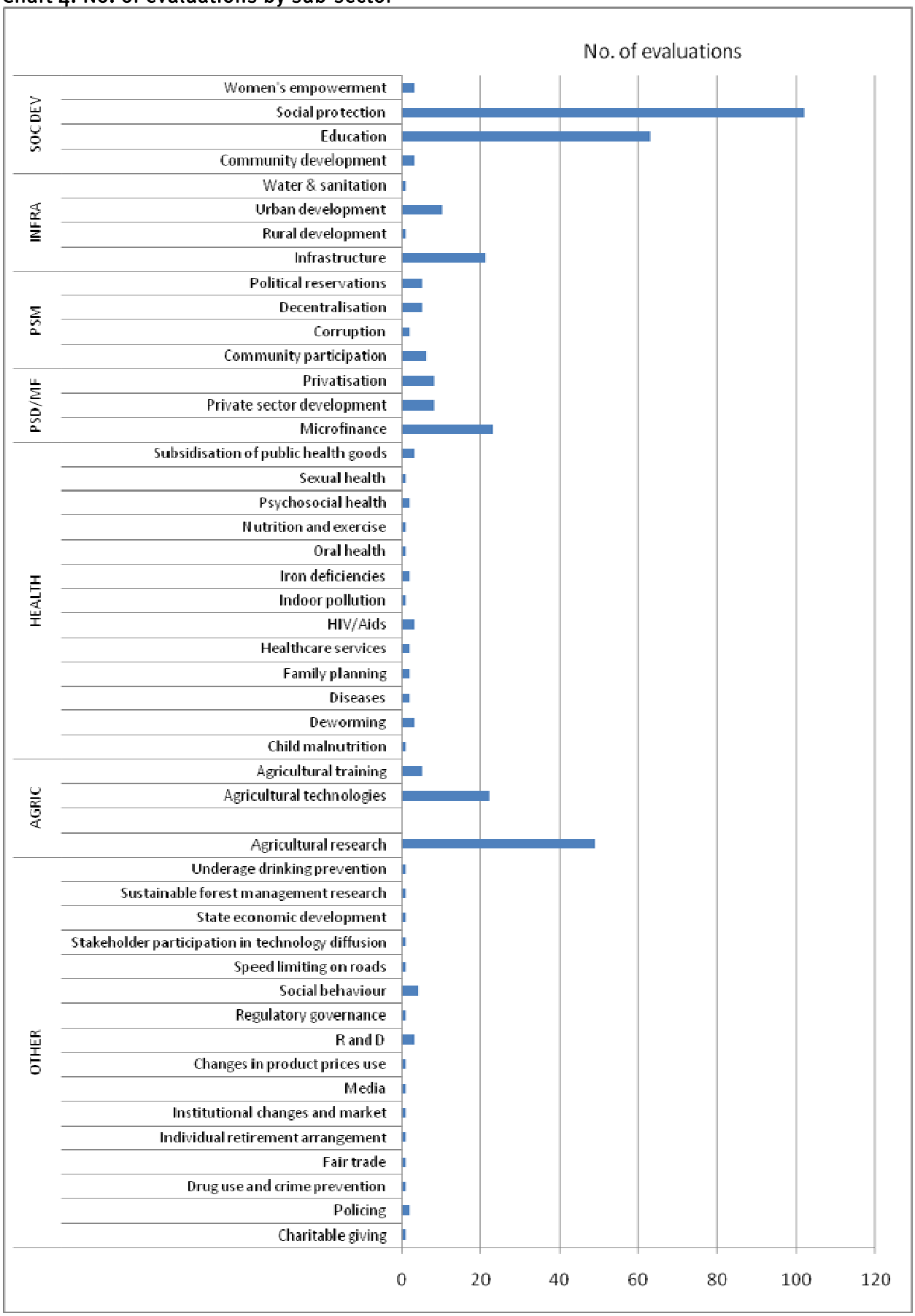
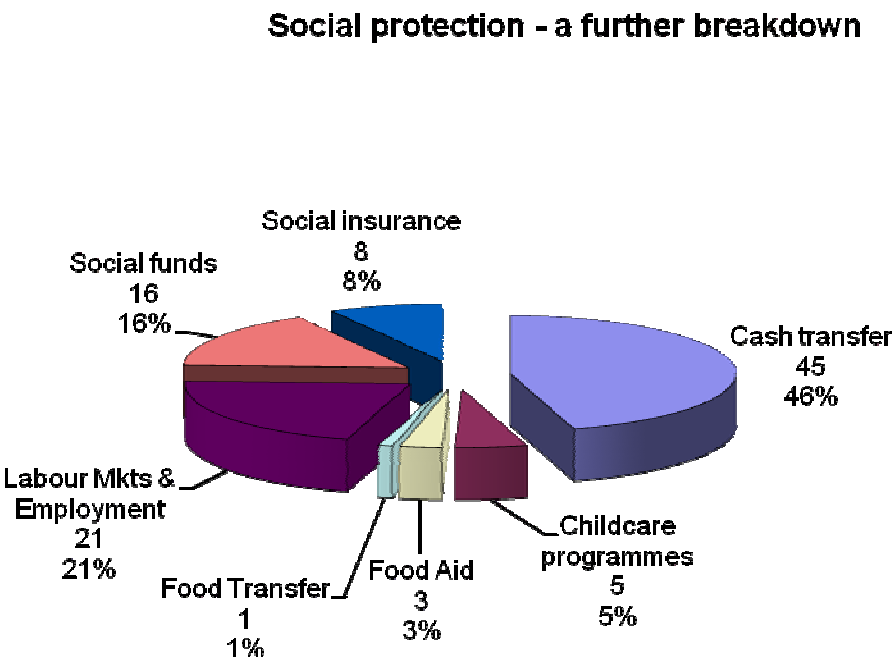
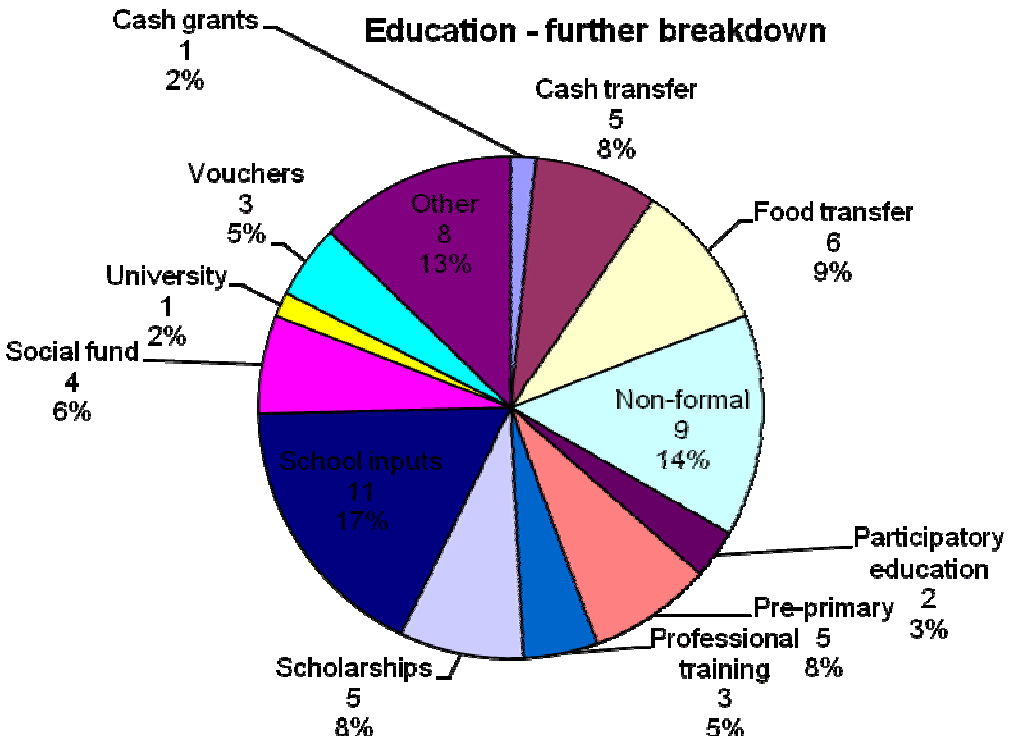


Chart 5: Social protection



As chart 7 illustrates, within the Education subcategory, the majority of IEs relate to school inputs (17%), followed by studies assessing the effects of Cash Transfers and School Vouchers[22] on a number of educational outcomes (13%). Studies analyzing the impact of non-formal education account for 14% of IE studies, followed by IEs studying the impact of various other education programmes (13%)

Chart 6: Education - a breakdown

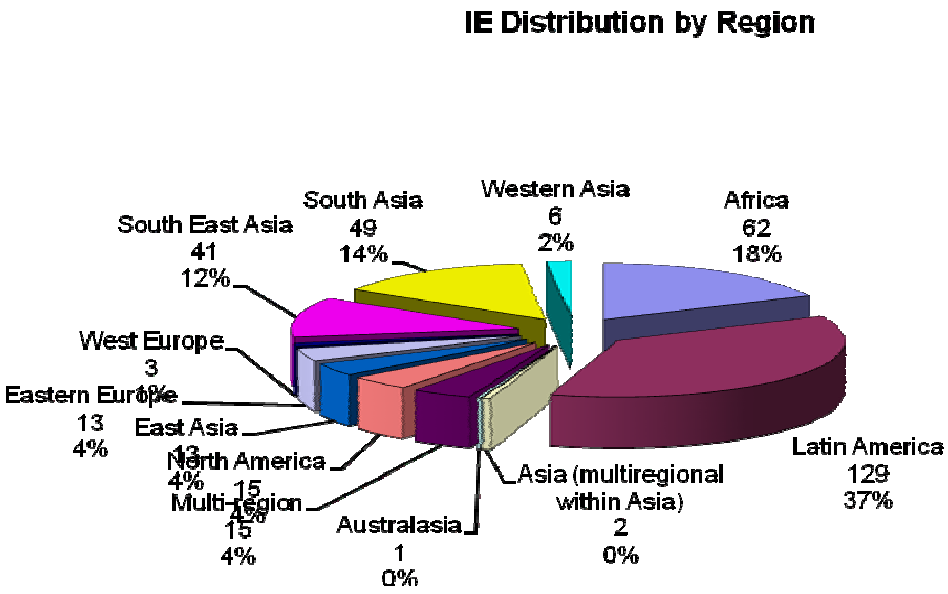


[22] As opposed to Cash and Food transfers within the Social Protection subcategory - which generally assess programmes in relation to a heterogeneous set of indicators (health-schooling-poverty) - those wihtin the Education subcategory focus specifically on educational outcomes.

Geographical coverage

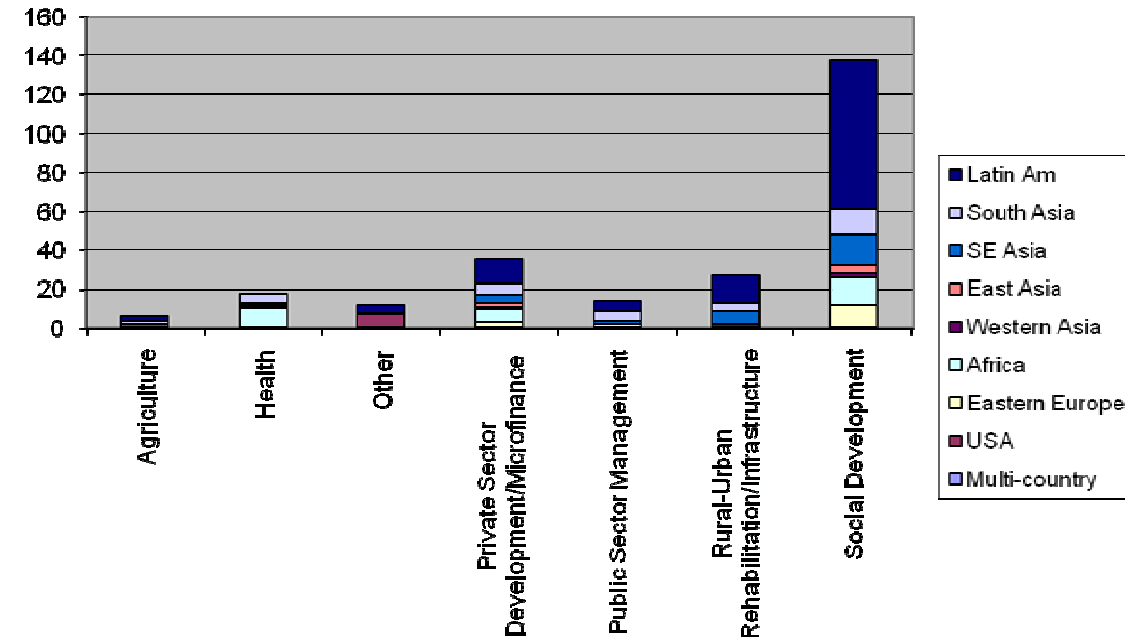
In terms of geographical coverage, the highest concentration of IEs can be found in Latin America *(see chart 8)*, and particularly in the sectors of Social Development, Rural-Urban Infrastructure and Private Sector Development/Microfinance. South Asia and South East Asia jointly add up to 26% of the total, while Africa accounts for 18%, followed by Eastern Europe (4%). China represents only 3% of the IEs in our database.

Chart 7: IE distribution by region



In terms of Sector-Region breakdown (see chart 9), despite the strong focus on Sub-Saharan Africa by the international development community, the thematic focus of IEs in Africa is also relatively narrow. Health represents approximately 50% of the total number of studies in the region, followed by Social Development, Private Sector Development/Microfinance and Public Sector Management. China's representation in our database is similarly narrow, relating to Social Development and Private Sector Development/Microfinance. Within the former, 2 out of the 5 IEs assess the impact of China's rural health insurance program (Social Protection), while the two microfinance studies evaluate the Southwest China Poverty Reduction Project.

Chart 8: Geographical breakdown of IE by sector



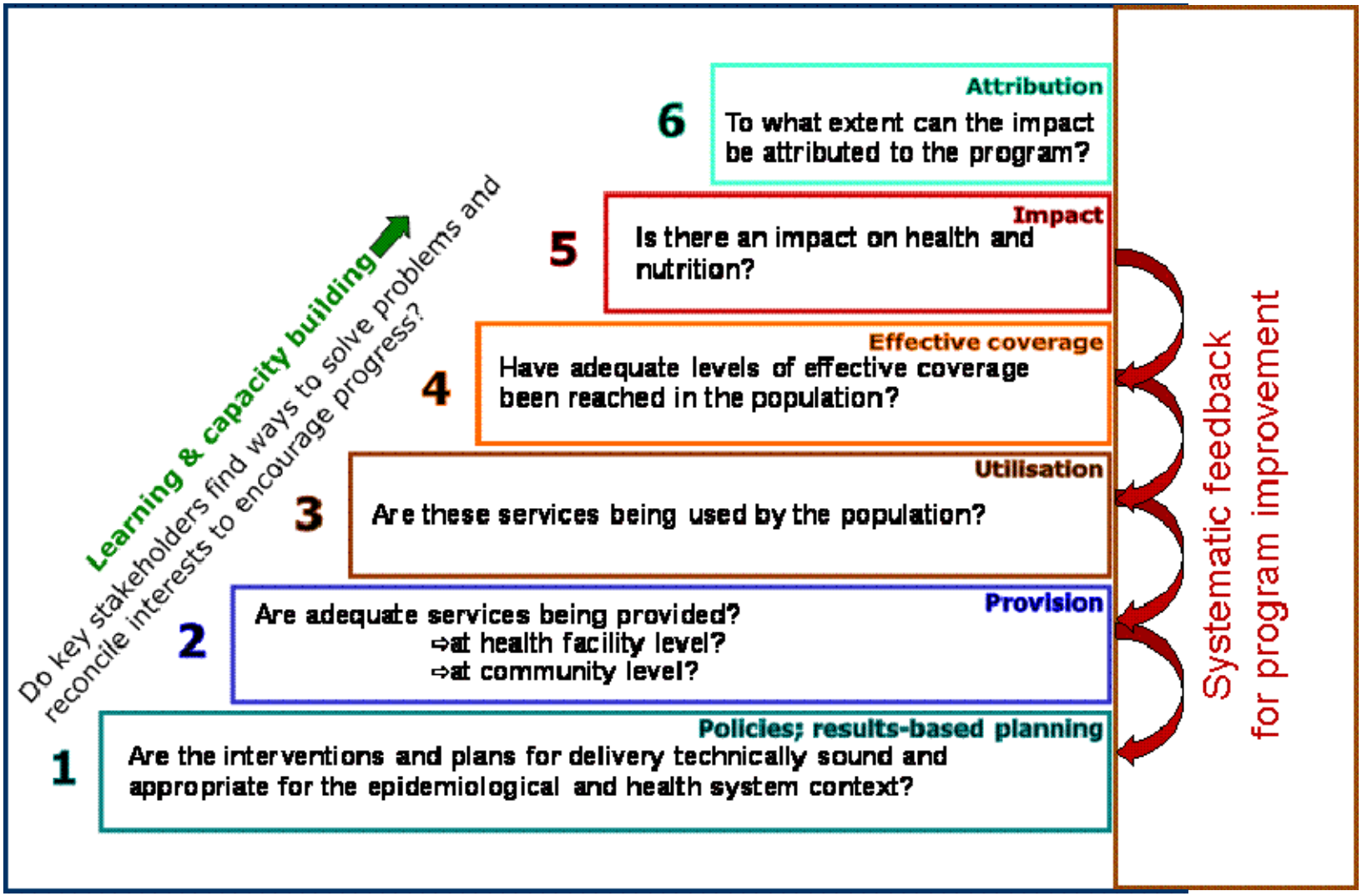
Appendix 2: Media coverage of impact evaluation findings

Source	Summary
Niranjan Rajadhyaksha (July 22nd 2008). 'Computers or Classrooms?' at: http://www.livemint.com/2008/07/22221226/Computers-or-classrooms.html?h=B	Fears erupted a decade ago about a growing 'digital divide' between the digital haves and have-nots. Policy makers have since explored ways this gap could potentially be bridged. Providing children in poor families' access to computers appeared an obvious first step. The One Laptop Per Child (OLPC) programme, for example, was sponsored by leading firms such as Google. Doubts have however been cast. The author draws on research from the Poverty Action Lab which explored whether academic performance improved with access to computers in Gujarat. Children in the schools assessed, in the slums of Ahmedabed and some other towns and villages, were given one hour at a computer daily. During this time, the teacher's role was restricted to switching the computers on and off. Leigh Linden at the Poverty Action Lab found that providing computers at schools is not much of an answer. A lot depends on how exactly they are used – as a complement or substitute to the teacher. 'The programme of computerized learning does not work too well when it is used to substitute the teacher in the normal school day. Math scores actually dropped in schools that took this path'. Lessons have also been drawn from a voucher scheme in Romania where some poor families received 200 Euros to purchase computers for their children. In the families who received the vouchers, it was found that children watched less television and spent less time on their homework compared with those who didn't (where this was the only difference between them). This is of importance given that schools need reforming. Linden points to more cost-effective ways to improve the academic performance of children from poor families. These include cash incentives for teachers, scholarships for girls, access to textbooks and good libraries. 'Computers are part of the answer — but perhaps not the most important part'.
Laura Vanderkam (July 1st 2008). 'Looking for the Virtuous Circle' at: http://www.american.com/archive/2008/may-june-magazine-contents/looking-for-the-virtuous-cycle	This article outlines the work of Ester Duflo at the Poverty Action Lab. In combating extreme poverty, she states 'unfortunately, we don't know very much about what works and what doesn't'. Duflo's research attempts to discover 'on the most micro level, how people make decisions, and how poverty changes the way people make decisions'. This is done through randomised trials which aim to understand the interplay of a multitude of factors, as done with the Education for All Programme in Kenya. In creating such trials Duflo works with NGOs and cooperative governments. Explaining the benefits of randomised trials Duflo states 'it's the truth, or as close as one can come to it in our messy universe'. A study by Duflo and Banerjee (2006) in Udaipur explored the microfinance revolution. Salaried jobs were compared to the occasional *dosa* (rice and bean pancakes) sale. It was concluded that 'The single most important characteristic of the middle class seems to be that they are more likely to be holding a steady job'. It's not that microfinance is misguided, it 'plays a role in helping the poor live a somewhat

	better life.' But a key difference between a solo *dosa* stand, and a packaged *dosa* factory that employs thousands of full-time workers across India, is more capital than any microfinance program is likely to provide. This highlights that effective aid must not only secure small businesses and intermittent labour, it must also help larger businesses flourish.
Sharad Raghavan (July 23rd 2008). 'Lessons From Copenhagen. Eighty per cent of the world's 140 million undernourished children lack essential micronutrients' at http://www.livemint.com/2008/06/23220407/Lessons-from-Copenhagen.html	The Copenhagen Consensus Conference concluded that the most important problem which needs to be addressed is malnutrition among children, particularly providing access to micronutrients such as vitamin A and zinc. The author holds that the Indian government should take the initiative in tackling undernourishment, particularly as India has more than one third of the world's malnourished children. The government's latest budget allocated less than 2% of its total plan expenditure to the development of women and children. A randomised study conducted by the Poverty Action Lab highlights this importance. It was found that providing iron supplements and deworming drugs to children throughout a preschool network led to significant weight gains. Preschool attendance also rose. 'Looking at the overall cost of the project, less than $2 per child per year, it is evident that such investment will yield great results'.
June 12th 2008 (No author) 'Proof that democracy Works? Health Services and Community-Based Monitoring in Uganda'. At: http://internationalbudget.wordpress.com/2008/06/12/proof-that-democracy-works/ Through the PAL website	This article draws on research carried out by Martina Bjorkman and Jakob Svensson for the Centre for Economic Policy Research which explored the impact of community based monitoring on the quality and quantity in health services. A Citizen Report Card methodology provided community members the opportunity to record their experiences and preferences of health services. Whether recommendations and desires were implemented was also recorded. This is thought to have provided incentives for improved services. The project was designed by staff at Stolkholm University, the World Bank and was implemented in cooperation with Ugandan organisations and practitioners. One year into the program, Bjorkman and Svensson found large increases in utilization, weight-for-age gains of infants, and markedly lower deaths among children. As such, it has been concluded that community monitoring can play an important role when top-down supervision proves ineffective. 'Macro-level research by political scientists has underlined the importance of the so-called 'democratic dividend'. While the link between democracy and concrete benefits to citizens can seem tenuous on a large scale, this project demonstrates that the links are much clearer at a local level.'
'Control Freaks. Are 'Randomised Evaluations' a Better Way of Doing Aid and Development Policy' *The Economist.* At: http://www.povertyactionlab.org/news/control%20freak.pdf June 12th 2008	This article discusses the rise in popularity of randomised trials by a group of economists at Harvard University and the Massachusetts Institute of Technology (MIT). Through this methodology, different policies are tested by assigning them to different groups. A celebrated example is throughout 20 antenatal clinics in Western Kenya where it was concluded that free distribution of anti-malaria bed nets is far more effective than charging even a small fee. The influence of randomised trials is growing. Last year the Spanish government gave the World Bank $16m to spend on evaluating projects in this way. It is questioned whether such evaluations are what they are cracked up to be. 'Randomistas' recently agreed that randomised evaluations are a good way to answer microeconomic questions, such as how to

	get girls to go to school, but tell us little about macro questions like budget policy. Advocates of the method, such as Banerjee, stand by the claim that 'the beauty of randomised evaluations is that the results are what they are', that they provide hard evidence. The method should be more widely applied. However, a tension appears to lie in the fact that 'policymakers do not want to know whether something works in a few villages. They want to know whether it will work nationwide. Here, randomised trials may not be quite so helpful.' Dani Rodrik has worried that the differences between randomistas and other economists risks re-opening a split between macro- and micro- economists which was beginning to close.
Michael Kremer (February 20th 2008). 'The Wisest Investment We Can Make: Using Schools to Fight Neglected Diseases'. *Global Health Policy.* At: http://blogs.cgdev.org/globalhealth/2008/02/the_wisest_investmen_1.php	Kremer makes the case that investing in deworming programmes throughout schools in developing countries has a great impact. This is backed by rigorous evidence including his study with colleague Edward Miguel. This evaluated the impact of the ICS school based treatment programme in Kenya. Treatment cut absenteeism by 25%. This evidence is also supported by Hoyt Bleakley at the University of Chicago Graduate School of Business who analysed the impact of a Rockefeller funded programme that treated worms in the US South at the beginning of the 20th Century. Here, school attendance also rose. Such programmes are also cost-effective. The Bush Administration and other institutions such as the World Health Organisation, the World Bank and groups like Partnership for Child Development are responding to the evidence. More Ministries of Education are beginning to take note.
Ijaz Kakakhel (October 31st 2008). 'Govt to include gender aspect in PC-1 formulation process.' At: http://www.dailytimes.com.pk/default.asp?page=2008%5C10%5C31%5Cstory_31-10-2008_pg5_14	The government in Pakistan is considering incorporating a gender impact variable into its project formulation process. Gender equality is held to be essential for national progress and development. Secretary Suhail Safdar, for instance, holds that such gender mainstreaming stands in line with the broader policy initiatives conceived in the Medium Term Development Framework and the Gender Reform Action Plans of the government. A more thorough gender perspective is needed throughout development projects and policies. Training workshops will be set up with the main aim of enhancing the capacity of government officials to mainstream gender in the formulation, implementation, monitoring and evaluation of government plans and policies.
Pan Yau (November 22nd 2007). 'Working Toward a Better Environment'. *China Daily.* At: http://www.chinadaily.com.cn/opinion/2007-11/22/content_6271575.htm	Many factors are to blame for the China's serious environmental problems. This article points to the lack of an environmental monitoring and evaluation system. The development of such a system is critically needed. The main resistance to environmental evaluation is seen to be a conflict of interests. What an evaluation system stresses is long term change which conflicts with the interests of different government bodies. A strategic system of environmental evaluation is essential if the concept of sustainable development is to be put into practice. Communication and coordination also need to be enhanced between different departments, to actively promote legislation on environmental legislation.

Appendix 3: Stepwise evaluation model



Appendix 4: Key informants

No.	Name	Affiliation	Type of Actor	Country	When consulted	Sector (if applicable)
1.	Howard White	World Bank EVD	Multilateral Organisation	USA	First study	
2.	Markus Goldstein	World Bank Poverty Reduction Group Economist	Multilateral Organisation	USA	First study	
3.	Judy L. Baker	World Bank	Multilateral Organisation	USA	First study	
4.	Michael Bamberger	World Bank, post-retirement consultant	Multilateral Organisation	USA	First study	
5.	Eduardo Masset	ex-World Bank EVD	Multilateral Organisation	USA	First study	
6.	Ole Winckler	DANIDA	Bilateral donor	Denmark	First study	
7.	Rachel Glennerster	Abdul Latif Jameel Poverty Action Lab, DFID Advisory Committee on Development Impact	Research Organisation	US	First study	
8.	Michael Quinn Patton	Independent Consultant (Organizational Development and Program Evaluation), Former President of the American Evaluation Association	Consultant	US	First study	
9.	Ruth Levine	Centre for Global Development	Research Organisation	US	First study	
10.	Ray Pawson	Realist Evaluation Specialist	Consultant		First study	
11.	Roger Riddell	Former International Director of Christian Aid; author of: "Does Foreign Aid Really Work?"	Civil Society Organisation	UK	First study	
12.	David Peretz	Independent consultant, IMF Evaluation Office	Multilateral Organisation	US	First study	
13.	Rick Davies	MANDE	Coordinator	UK	First study	
14.	Juliet Pierce	Performance Review Assessment Centre (PARC)	Research Organisation	UK	First study	
15.	David Raitzer	Centre for International Forestry Research (CIFOR)	Research Impact Assessment Scientist	Indonesia	First study	
16.	David Lewis	London School of Economics Dept of Social Policy and Centre for Civil Society	Academic organisation	UK	First study	
17.	John Lavis, MD	McMasters University; Director of Program in Policy Decision	Academic organisation	Canada	First study	

No.	Name	Affiliation	Type of Actor	Country	When consulted	Sector (if applicable)
		Making				
18.	Zenda Ofir	African Evaluation Association (AfrEA)	Association	South Africa	First study	
19.	Javier Escobal	Group for the Analysis of Development (GRADE)	Research Organisation	Peru	First study	
20.	Priyanthi Fernando	Centre for Poverty Analysis (CEPA)	Research Organisation	Sri Lanka	First study	
21.	Norma Correa Aste	Economic and Social Research Consortium (CIES)	Research Organisation	Peru	First study	
22.	Pak Sudarno	SMERU Research Institute	Research Organisation	Indonesia	First study	
23.	Prof S. Galab	Centre for Economic and Social Studies	Research Organisation	Hyderabad, Andhra Pradesh, India	First study	
24.	Yoshio Wada	National Graduate Institute for Policy Studies	Academia	Japan	Second Study	Infrastructure and rural/urban development
25.	Jonathan Zinman	Department for Economics, Dartmouth College	Academia	US	Second Study	Private sector/ microfinance
26.	Victoria Elliot	Consultant, former unit manager, corporate evaluation and methods, IEG, World Bank	Consultant	US	Second Study	Economic development
27.	Jan Isaksen	Chr Michelsen Institute	Research Organisation	Denmark	Second Study	Infrastructure and rural/urban development
28.	Maximo Torero	Division Director, Markets, Trade and Institutions, IFPRI	International Research Organisation	US	Second Study	Infrastructure and rural/urban development
29.	Luis Teodero Marcano	Inter-American Development Bank	Multilateral Organisation	US	Second Study	Infrastructure and rural/urban development
30.	Stephen Brushett	Infrastructure specialist, World Bank	Multilateral Organisation	US	Second Study	Infrastructure and rural/urban development
31.	Sheelagh O'Reilly	Research into Use programme	Research intermediary	UK	Second Study	Renewable and Natural Resources
32.	Ruth Meinzen-Dick	Senior Research Fellow, IFPRI	Researcher	US	Second Study	Renewable and Natural Resources
33.	Ade Freeman	Targeting and innovation director, ILRI		Kenya	Second Study	Renewable and Natural Resources
34.	Doug Gollin	Associate Professor, Department of Economics, Williams Colleage	Academia	US	Second Study	Renewable and Natural Resources

No.	Name	Affiliation	Type of Actor	Country	When consulted	Sector (if applicable)
35.	Doug Horton	International Service for National Agricultural Research (ISNAR)	International Research Organisation	Netherlands	Second Study	Renewable and Natural Resources
36.	Jock Anderson	Emeritus Professor, Faculty of The Professions, School of Business Economics and Public Policy, University of New England	Academia	US	Second Study	Renewable and Natural Resources
37.	Derek Byerlee	Former World Bank, now independent consultant		US	Second Study	Renewable and Natural Resources
38.	Debbie Templeton	Impact Assessment Program Manager, Australian Centre for International Agricultural Research (ACIAR)	Donor	Australia	Second Study	Renewable and Natural Resources
39.	Hoa Ngo Thi Quynh	Senior Programme Officer, DFID	Bilateral Donor	Vietnam	Second Study	Infrastructure and rural/urban development
40.	Karen Proudlock	Research Officer, ALNAP	Intermediary	UK	Second study	Humanitarian
41.	John Mitchell	Research Fellow, ALNAP	Intermediary	UK	Second study	Humanitarian
42.	Charles-Antoine Hofmann	Humanitarian Policy Adviser, British Red Cross	Policy adviser	UK	Second study	Humanitarian
43.	Jodi Nelson	Director of Research & Evaluation, International Rescue Committee	Evaluation	US	Second study	Humanitarian
44.	Peter Walker	Irwin H. Rosenberg Professor of Nutrition and Human Security Director, Feinstein International Center, Tufts University	Academia	US	Second study	Humanitarian
45.	Antonella Mancini	Independent consultant	Consultant	UK	Second Study	Humanitarian
46.	Paul Glewwe	Professor, Department of Applied Economics, University of Minnesota	Academia	US	Second study	Social Development
47.	Karthik Muralidharan	Professor, Department of Economics, University of California	Academia	US	Second study	Social Development
48.	Samuel Berlinski	Lecturer, Department of Economics, University College London	Academia	UK	Second study	Social Development
49.	Rachel Glennester	Department of Economics, MIT	Research Organisation	US	Second study	Social Development
50.	Akter Ahmed	Senior Research	International	IFPRI	Second study	Social

No.	Name	Affiliation	Type of Actor	Country	When consulted	Sector (if applicable)
		Fellow, Food Consumption and Nutrition	Research Organisation			Development
51.	Gonzalo Hernandez	Director of Research, CONEVAL, Mexico	Government	South	Second study	Social Development
52.	Laura Rawlings	Country Sector Leader for Central America in the Latin America and Caribbean Human Development Department at the World Bank	Multilateral Organisation		Second study	Social Development
53.	Ruth Levine	Vice President for Programs and Operations, and Senior Fellow, Centre for Global Development	Research Organisation	USA	Second study	Health and Results Based Aid
54.	Paul Bolton	Associate Scientist, Center for Refugee and Disaster Studies Johns Hopkins Bloomberg School of Public Health	Academia	USA	Second study	Health
55.	Jennifer Bryce	Senior Scientist, John Hopkins Bloomberg School of Public Health	Academia	USA	Second study	Health
56.	Charlie Teller	Bixby Visiting Scholar Population Reference Bureau	Bilateral donor	USA	Second study	Health
57.	Owen Barder	DFID	Bilateral donor	UK	Second study	Results Based Aid
58.	Jeremy Clarke	DFID, retired	Bilateral donor	UK	Second study	Results Based Aid
59.	Chris Adam	Oxford University	Academia	UK	Second study	Results Based Aid
60.	Jan Willem Gunning	Free University Amsterdam	Academia	Netherlands	Second study	Results Based Aid
61.	Lars Johannes	World Bank GPOBA	Multilateral Organisation	USA	Second study	Results Based Aid
62.	Andrew Lawson	Fiscus	Research Organisation	UK	Second study	Results Based Aid
63.	Bill Savedoff	Social Insight	Research Organisation	USA	Second Study	Results Based Aid