ALINe
PEOPLE-CENTRED PERFORMANCE

Working Paper 8

Evaluation in Theory and Practice

Henry Lucas

May 2012

About ALINe

The Agricultural Learning and Impacts Network (ALINe) is an initiative between the Institute of Development Studies, UK and other partners, with funding from the Bill & Melinda Gates Foundation (BMGF). ALINe combines practical experience and technical expertise with cutting-edge academic research in order to learn from farmer experience, share good practice and pilot new approaches to monitoring and evaluation. ALINe utilises a people-centred performance measurement that is responsive to poor farmers, offering the potential to re-incentivise and re-align the aid system. ALINe has multiple clients in the public, private and third sectors and a portfolio currently in 9 countries with emphasis in Sub-Saharan Africa and South Asia. Visit our website: www.aline.org.uk

About IDS

The Institute of Development Studies is one of the world's leading charities for research, teaching and communications on international development. Founded in 1966, the Institute enjoys an international reputation based on the quality of its work and the rigour with which it applies academic skills to real world challenges. Its purpose is to understand and explain the world, and to try to change it – to influence as well as to inform.

IDS hosts five dynamic research programmes, five popular postgraduate courses, and a family of world-class web-based knowledge services. These three spheres are integrated in a unique combination – as a development knowledge hub, IDS is connected into and is a convenor of networks throughout the world.

The Institute is home to approximately 80 researchers, 50 knowledge services staff, 50 support staff and about 150 students at any one time. But the IDS community extends far beyond, encompassing an extensive network of partners, former staff and students across the development community worldwide. Visit our website: www.ids.ac.uk

Evaluation in Theory and Practice
Henry Lucas

ALINe Working Paper

Available from:
Communications Unit, Institute of Development Studies, Brighton BN1 9RE, UK
Tel: +44 (0) 1273 915637
Fax: +44 (0) 1273 621202
E-mail: bookshop@ids.ac.uk
Web: www.ids.ac.uk/ids/bookshop

IDS is a charitable company limited by guarantee and registered in England (No. 877338)

Evaluation in Theory and Practice

Henry Lucas

Henry Lucas is a Fellow of the Institute of Development Studies. A statistician by background, he has specialised in information systems, M&E and research methods, primarily in the health sector. Current research includes work on social protection for the Future Health Systems Consortium and links between poverty and major illness in Cambodia, China and Lao. He has worked extensively on PRSP monitoring, undertaking several general reviews and developing monitoring frameworks for Sierra Leone and The Gambia. He is adviser on M&E to a DFID/NORAD MCH programme in Northern Nigeria and has worked for many years on health M&E in China.

Acknowledgements

The author would like to thank Yvonne Pinto and Johanna Lindström for their inputs to this paper.

Contents

Lists Abbreviations, Tables, Figures and Boxes

Abbreviations

ADB	Asian Development Bank
ALINe	Agricultural Learning and Impacts Network
CIOMS	Council for International Organizations of Medical Sciences
CMO	Context + Mechanism = Outcome
CoIMPact	Consultative impact monitoring
CRCT	Clustered randomised controlled trial
CWIQ	Core welfare indicator questionnaire
DALY	Disability-adjusted life year
EC	European Commission
IMF	International Monetary Fund
LFA	logical framework approach
MDG(s)	Millennium Development Goal(s)
M&E	Monitoring and Evaluation
OECD	Organisation for Economic Co-operation and Development
PAF	Performance assessment framework
PBM	Performance based monitoring
PETS	Public expenditure tracking survey
PIA	Poverty impact assessment
PM&E	Participatory monitoring and evaluation
PPA	Participatory poverty assessment
PRS	Poverty reduction strategies
PRSP	Poverty reduction strategy paper
PSIA	Poverty and social impact analysis
QIM	Qualitative impact monitoring
QSDS	Quantitative service delivery survey
RBME	Results based monitoring and evaluation
RCT	Randomised controlled trial
RE	Realistic evaluation
TBE	Theory based evaluation
ToC	Theories of Change

Lists of Tables, Figures and Boxes

Summary

Monitoring and evaluation systems can deliver essential information not only to guide the implementation of an intervention and determine if it should be regarded as having succeeded or failed, but opportunities to learn why and under what conditions similar outcomes would be more or less likely if the intervention were repeated. However, there is a growing consensus that the unconsidered application of traditional, project-oriented approaches to M&E often fail to deliver expected benefits. This is usually attributed to poor execution. All too often there is a failure to integrate M&E into intervention planning and implementation processes, limited capacity among those responsible for M&E, limited understanding of its potential value among other staff (M&E primarily seen as an additional burden) and insufficient resources to deliver findings of an appropriate quality. However, in many cases limited thought will also have been given to the theoretical framework which, explicitly or implicitly, informs the design of an M&E system and the implications of that framework in terms of the range of evaluation objectives that may be prioritised by different stakeholders. For example, many observers complain of a 'compliance culture', expressed in M&E systems that prioritise accountability to those funding an intervention, often at the cost of equally important objectives, for example those relating to stakeholder engagement and systematic learning.

In response to such concerns, this working paper discusses both current debates relating to evaluation theory and some recent examples of the practice of M&E in a series of Poverty Reduction Strategies. The design of these strategies, typically supported by the World Bank, IMF and other international agencies, has placed considerable emphasis on the delivery of useful and timely performance indicators, often using innovative approaches. Using examples from a range of sectors, but with a particular focus on health interventions, the author then considers their relevance for agriculture, while recognising that this sector many have intrinsic features that make the design of effective M&E systems even more problematic. Three competing perspectives on evaluation – experimental, theory based and realist – are reviewed and analysed. One key observation of the paper is that a clear statement of priorities is an essential step in both the choice of evaluation approach and the detailed design of M&E systems. Appropriate evaluation strategies may vary substantially depending on the priorities attached to a range of objectives including: planning/efficiency, accountability, implementation, knowledge production, or institutional and network strengthening. To conclude, the author explores the practicality of implementing a 'combined methods' approach, with the potential to satisfy multiple objectives.

Introduction

The characterisation of the existing state of M&E in agricultural interventions by those participating in the ALINe consultation survey (Lindstrom 2009; Haddad *et al.* 2010) may seem familiar to many working in other areas: a 'compliance culture', expressed in a preoccupation with accountability to donors rather than to intended beneficiaries and other stakeholders; failure to fully integrate M&E into intervention planning and implementation processes; limited capacity among those responsible for M&E; limited understanding of its potential value among other staff (M&E primarily seen as an additional burden); and insufficient resources to deliver findings of an appropriate quality. These are common complaints, with underlying causes linked to deeply entrenched attitudes that either attach limited importance to accountability and transparency or are reluctant to allocate the often substantial resources required to achieve them. To some extent they probably also reflect a failure on the part of the evaluation community, either in terms of providing convincing evidence of the value of their activities or in finding effective methods to promote them.

However, research by ALINe suggests that agricultural interventions may have intrinsic characteristics that make particular demands on M&E (see Haddad *et al.* 2010). These include:

- a lack of clarity as to primary objectives – projects typically have multiple objectives entailing complex tradeoffs;
- long 'causal chains', in terms of both number of links and overall project duration (Millstone *et al.* 2010; and
- sensitivity to uncertainties imposed by climate and other natural phenomena, accentuating the potential disconnect between individual incentives and programme impacts (Sabates-Wheeler *et al.* 2010).

The resulting difficulty in specifying the 'implementation theory' (Weiss, 1995) of such interventions is seen as seriously impeding the design of an appropriate M&E system. In the absence of a realistic model of the process by which an agricultural intervention is intended to translate inputs into clearly identified outcomes, it is very difficult to know how to monitor or evaluate performance.

Is the position substantially worse than in other sectors? This paper considers both theoretical debates and selected practical applications in order to draw potential lessons for the evaluation of agricultural development interventions. It is based on a necessarily limited review of the very substantial literature on evaluation theory but does attempt to explore material beyond the development studies area, with a particular focus on the leading English-language European journals. These were seen as a potentially very interesting source because of the diverse political and socio-economic contexts in which evaluations were designed and implemented. The discussion of practice is based on the wide range of monitoring and evaluation activities associated with implementation of the World Bank/IMF 'Poverty Reduction Strategies' (PRS). In this case the choice of source material was based on an awareness of the extent of innovatory monitoring practices resulting from the emphasis on the routine estimation of timely, reliable indicators as part of the annual PRS assessment procedures.

One characteristic of the European evaluation material which became evident at an early stage of the review was that innovation was mainly in terms of methodology and not of methods. This almost certainly reflects the data-rich environment within which these evaluations are undertaken. For example, most medium to large scale public and private sector European institutions have sophisticated management information systems that can be used to access detailed, highly reliable time-series data on areas including finance, human resources and operational activities. Many adopt 'performance-based management systems' that generate a wide range of key performance indicators on a routine basis. Within this context evaluators rarely have to devise the type of innovative methods which are often essential in many less developed countries. Their basic strategy therefore typically involves document review, analysis of existing routine data systems, key informant interviews and focus groups discussions.

This review examines some of the current theoretical debates using this source and other relevant material in section 1. The second section focuses on the practical implementation of monitoring and evaluation frameworks developed for Poverty Reduction Strategies (PRS), where resource constraints mean that discussion of innovative methods is very much to the fore. The paper concludes with reflections on lessons for M&E in agriculture, particularly considering how to combine different methods and approaches to support more effective M&E.

1 Developments in evaluation theory

1.1 A variety of perspectives on evaluation

Debates on evaluation methodology are not unique to the international development sector. Over recent years, public policy reforms, including out-sourcing, public-private partnerships, increased marketisation of social services and a focus on 'results based management', have had radical implications for those working in evaluation across sectors in OECD countries. Europe is of particular interest because of the several major policy reforms and initiatives (related to European Union expansion) that have been implemented with a variety of approaches in a relatively short time period. This reflects the cultural diversity and the widely varying capacity and perceived role of the state across the European nations (Stern 2004). This led to a substantial and rapidly expanding literature which raises questions not only on evaluation methodology, but also on the underlying rationale and objectives of different evaluation paradigms.

The diversity of this literature is illustrated by Stern (2009: 5) who describes the 'daunting task of comparing and assessing the merits of what is being advocated" and indicated the value "for the perplexed reader (including this Editor) when authors offer articles that compare or even integrate different approaches'.

Many of the core issues are close to those raised in the development literature. In an earlier editorial, Stern highlighted an:

- increasing focus on "'theory'-based approaches and a renewed interest in methods and the nature of evidence";
- increased resistance to the assumption that an evaluation should primarily aim to determine the overall success or failure of an intervention;
- increased diversity of backgrounds, specialisations and approaches of people undertaking evaluations;
- the need to 'achieve coherence between sometimes incompatible evaluation objectives and questions';
- a desire for ever closer 'engagement with civil society – including communities, the private sector and public service users';
- the need to move beyond single evaluations to 'synthesis reviews and meta-analyses' (Stern 2008).

Several authors see these developments as related to the relatively limited role evaluation has played in influencing public policy. Many agree the developments in Europe are a result of government proclamations of its commitment to 'evidence-based policy' and 'results-based management' in the public sector. But, there is a growing concern that conventional evaluation approaches have not delivered evidence that can be readily absorbed into existing policy and management decision-making processes. Many sense '… a growing disillusionment with conventional evaluation praxis. Many governments experience only limited use of evaluation findings. Evaluation findings do not automatically feed back into a receptive and responsive decision-making process' (Bastoe 2006: 97).

The European Commission's (EC) guide to evaluating socio-economic development initiatives suggests that one cause of these concerns relates to 'incompatible evaluation objectives and questions'. The guide describes five purposes which stakeholders may prioritise for an evaluation (European Commission 2007):

1. Planning/efficiency: ensuring that there a policy/programme is justified, and that resources are efficiently deployed.
2. Accountability: demonstrating how far a programme has achieved its objectives and how well it has used its resources.
3. Implementation: improving programme performance and effectiveness in terms of delivery and management.
4. Knowledge production: increasing understanding of what works and when, and how interventions can be made more effective.
5. Institutional and network strengthening: improving and developing capacity among programme participants and their networks and institutions.[1]

These influence methodological preferences:

- Planning and efficiency issues may be seen as best approached through various forms of impact and cost-benefit analysis, linked to the traditional logical framework approach (LFA) (EuropAid 2004).

[1] These purposes are similar to ALINe's five purposes of M&E (see Haddad et al. 2010).

- Those most concerned with accountability will tend to focus on the assessment of performance against agreed targets and/or benchmarking against comparator interventions. The emphasis here is usually on quantitative techniques, including a conventional auditor-style analysis of monetary measures.
- Evaluators involved with implementation may promote the use of 'formative evaluation' methods that can provide rapid feedback on processes and interim outcomes for management purposes. This could be combined with institutional analysis to assess the performance of administrative and service delivery units to understand reasons for success or failure.
- The knowledge production agenda is seen as prioritising rigour, representativeness and the 'cautious interpretation of findings, especially where these may be inconsistent'. Two competing paradigms are identified: the 'experimental', based on the methodology of the controlled trials used to evaluate clinical treatments, and the 'realist', focused on case-studies that allow detailed comparative analysis of different 'intervention/outcome/context configurations'.
- Institution and network strengthening will be primarily concerned to ensure that evaluation is meeting the needs of all stakeholders and promoting their involvement and effectiveness in all aspects of the evaluation process.

A recent review of health sector interventions (Peters et al. 2009) similarly argues that different types of scientific evidence are required depending on the objectives of an evaluation. It identifies four types of judgment that policymakers may make about a health intervention (Table 1). To determine if an intervention has attained intended targets, for example population coverage, the type of inference described by Habicht et al. (1999) as 'adequacy' will be appropriate. Attribution is assumed in these cases and a simple before-and-after (or preferably time-series) study will suffice. If external factors may have confounded the relationship between intervention and outcomes, a 'plausibility' argument (Habicht et al. 1999; Victora et al. 2004) may be required. This implies a need for using a comparator group to construct a counterfactual – what would have happened without the intervention? Finally, for statistically valid, confidence limited estimates of differences between indicators of change for intervention and non-intervention sites, 'probability' inference based on randomised controlled trials (RCTs) is required.

Table 1: Judgments on health interventions and implications for evaluation design

Type of Judgment	Primary question to be answered	Type of inference	Evaluation design
Intervention is efficacious/effective	Is any measured effect on health services or health status attributable to the intervention?	Probability	Controlled trials, usually randomising clusters rather than individuals, intervention implemented in some areas and not others
Intervention is likely effective	Is any measured effect on health services or health status likely due to the strategy rather than other influences?	Plausibility	Concurrent, non-randomised clusters where intervention implemented compared to where not; before-after or cross-sectional study of intervention and non-intervention populations
Demonstration of expected changes in behaviours, health services or health status	Are behavioural, health services or health indicators changing among beneficiaries of an intervention?	Adequacy	Before-after or time-series in intervention population only
Explanation of how or why an intervention works	How did intervention lead to measured effects on health services or health status?	Explanatory	Repeated measurements of variables on context, actors, implementation depth and breadth across subunits. Key informant interviews, focus groups, historical reviews, and triangulation of data sources.

Source: adapted from Peters et al. 2009.

This health sector review notes that none of the first three designs will meet policymakers' requirements who wish to be told precisely how the intervention achieved its specified outcomes; this is related to the 'knowledge production' objective. The review suggests, 'Perhaps the most important lessons from this enquiry are not about what should be done to improve health services, but learning about how to use knowledge to improve health services'. Determining that a particular intervention can improve health outcomes is typically much easier than, for example, explaining "how to provide effective training and supportive supervision, ensuring that materials are

available to do the job, and that the right incentives and accountabilities for health workers to perform are in place" (Peters et al. 2009: 9).

Looking at the Milne et al. (2004: 339) classification (Table 2), the evaluation designs in Table 1 can seem as derived from a single paradigm, which regards the needs of 'decision-makers' as paramount. The 'pragmatic' or 'utilisation-focused' approach (Patton 1997) 'answers the question of whose values will frame the evaluation by working with clearly identified, primary intended users who have responsibility to apply evaluation findings'. The role of the evaluator is to help users to 'select the most appropriate content, model, methods, theory, and uses for their particular situation' (Patton 2002). This is in sharp contrast to the 'constructivist' or 'fourth generation' evaluation paradigm (Lay and Papadopoulos 2007 Guba and Lincoln 1989), which aims at a negotiated settlement between all stakeholders, attempting to reconcile their diverse perceptions. 'Experimental' evaluators are characterised as strict positivists who see their task as identifying cause and effect relationships using controlled trials, while those adopting the 'theories of change' approach (incorrectly identified with 'realists' as discussed below) insist on the need for theoretical explanations of those relationships. Finally, the 'pluralists' seek ways to draw on all these different perspectives and are usually condemned as unprincipled eclectics.

Table 2: Characteristics of different approaches to evaluation

Perspective	The theory	Approach
Experimental	A system of cause and effect is assumed to exist, which cannot be observed directly. Causation can only be inferred through controlled observations.	Randomised or quasi-experimental trials with pre-test, post-test, and control group.
Constructivist	Follows the idea that truth is always attached to some standpoint rather than being external to any one group.	Qualitative techniques used to explore meanings that stakeholders attach to phenomena, aiming to reconcile different meanings through a consensual process.
Pragmatic	Regards as valid knowledge that which is considered pragmatically acceptable by decision-makers.	Qualitative and quantitative techniques used to produce the evidence decision makers need.
Pluralist	Takes the view that knowledge produced from alternative perspectives all add important insights to events.	Qualitative and quantitative techniques are combined to gain greater insight into the working of an intervention and to help define the causal pathways that might exist.
Theories of change (realist)	Evaluations are built around explicit theories of how interventions work in specific contexts.	Qualitative and quantitative techniques used to test theories.

Source: adapted from Milne et al. 2004: 339.

1.2 Experimental versus theory based evaluation

The central debate in the evaluation literature concerns the relative merits of the experimental and two alternative, 'theory based', evaluation (TBE) paradigms: 'theories of change' and 'realist'. As noted in the previous section, the wide variety of theoretical concepts, frameworks and terminologies promoted by different authors and agencies makes the deconstruction of the evaluation debate far from simple. This debate as to which methodologies can best describe and attribute causality (and contribute to the knowledge production objective) in evaluating interventions has been called 'the causal wars' (Scriven, 2010; Stern, 2008).

The experimental perspective: randomised controlled trials

The European Commission Guide (2007) suggests an increasing disaffection amongst policymakers and practitioners with the 'orthodox' approach to evaluation that is based on the logic of scientific experimentation. The 'gold standard' under this approach is the randomised controlled trial. This involves randomly allocating members of a given population to one of two groups. Both groups should then be isolated from all potentially confounding external factors and only one of them subjected to an intervention. A further requirement, routinely applied in 'double-blinded' clinical health trials, is that neither the population nor those implementing the intervention should know who is receiving the treatment. The aim is to avoid a 'placebo' effect - where knowing that the treatment can be potentially beneficial may itself encourage a positive response, irrespective of any direct physiological effects. Comparing the two groups 'before and after' the intervention is then regarded as providing a direct measure of its impact. Given that random allocation and isolation from external factors are often impossible

in practice, the tasks of the evaluator are seen as (a) approximating these conditions to the extent possible (for example by careful 'matching' of the members of the control and treatment groups) and (b) allowing for any unavoidable divergences by careful analysis and interpretation (Shadish, Cook and Campbell 2001).
Though the application of this approach to complex policy interventions has long been contested, it retains many serious and influential advocates (e.g. Kramer and Holla 2008; Banerjee and Duflo 2008). The major collaborations on systematic reviews, Cochrane (Higgins and Green 2008) and Campbell (Davies and Boruch 2001) also tend to privilege experimental methods in general and RCTs in particular, often excluding studies which use alternative approaches. Indeed, Hansen and Rieper (2009), suggest that such policies, given the status enjoyed by these two organisations, have had major consequences in terms of framing international research priorities, encouraging researchers to focus on issues which can be more readily addressed by using this approach.

Opponents have typically argued on practical grounds (e.g. Chambers 2008). They suggest that is often impossible to attain the necessary conditions for implementing the above model; sometimes physically impossible but more commonly because of resource costs or violation of ethical standards involved in implementing required matching or control procedures. Common objections include:

- In practice, most interventions do not involve a single, well specified 'treatment' but a range of treatments, often tailored to specific circumstances of individual members of the treatment group. For example, a relatively simple education project might involve an initial allocation of basic equipment items to ensure that all schools in the treatment group conform to a minimum standard. The inputs required to achieve this aim may vary substantially from one school to another.

- Generally, interventions are almost never implemented 'as planned' but 'as modified' to conform to local realities – political, social, cultural, etc. – in each treatment site. For example, exemption schemes intended to reduce the cost of care for poor households will typically be implemented according to rules determined by local managers – not central government officials (e.g. Gilson et al. 2001).

- The practical reasons for using 'clustered RCTs' (CRCTs) where treatment and control groups have clusters of individual 'members' (villages, schools, hospitals, districts, etc.), often result in relatively small treatment and control groups. This restricts the scope for persuasive statistical analysis, especially where there is a need for post-hoc control of potentially confounding factors.

- The complexity of the contexts within which interventions are often undertaken, and the multiple factors that may influence outcomes, severely limits the possibilities for extrapolating findings to other locations, populations or time periods (Eldridge et al. 2008; Green and Glascow 2006). For example, a CRCT in the Mwanza district of Tanzania appeared to demonstrate conclusively that effective treatment of sexually transmitted infections could reduce the incidence of HIV; however, subsequent disappointing results from 'similar' exercises in other settings generated a series of plausible context-linked explanations (e.g. Grosskurth et al. 2000).

- Randomised allocation is rarely feasible and 'matching' of treatment and control groups, especially in the case of CRTCs is often highly problematic not only practically (in terms of required data) but also conceptually (White 2007). One often overlooked issue is that appropriate matching may require not only that groups are currently similar but that they have a similar history. For example, two communities may appear to have very similar socio-economic characteristics at a given point of time even if one has reached that point after a period of steadily growth and the other in consequence of progressive decline.
- Many interventions have lengthy timescales and it is often impossible to control or even adequately assess the impact of multiple external factors – other interventions, policy changes, social unrest, climate variations – which may affect control and treatment groups differently.

In a recent lecture, Angus Deaton (2009: 3), argued that 'in ideal circumstances, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow to tell us "what works" in development, to design policy, or to advance scientific knowledge about development processes. Project evaluation using randomized controlled trials is unlikely to discover the elusive keys to development, nor to be the basis for a cumulative research program that might progressively lead to a better understanding of development.' Many leading advocates of randomised experiments in social research are well aware of these criticisms, but argue that they are by no means restricted to their preferred methodology. 'If our theories are not very good, and the impact of treatment depends on context in a way that is complicated, subtle, and difficult to predict, results from one setting are unlikely to generalize in other settings that may look similar to reasonable people. If indeed it is so difficult to generalize, then

this would raise questions not simply about randomized evaluations but more generally about the extent we can learn from social science' (Kremer 2008: 3).

In recent years, the predominance of RCTs has been seriously challenged in its primary territory - the evaluation of clinical treatments of specific health conditions. A special issue of the Journal of Clinical Epidemiology (January 2009) discussed the potential role of 'pragmatic' trials. In this, Zwarenstein and Treweek (2009) expressed a concern shared by many practicing physicians - that the vast majority of clinical trials (all but 100 of 250,000 reviewed) could be described as 'explanatory' (designed to test a hypothesis in a highly controlled context), rather than 'practical' (designed to identify interventions that might produce beneficial outcomes in practice). They argue that by adopting 'laboratory' conditions (Table 3), and excluding patients with conditions that might confound the relationship (between treatment and outcome), clinical trials produce findings which may be of scientific interest but are of limited practical value to clinicians working 'in the real world' and having to make decisions on best practice.

The authors point out that the focus on explanatory trials could be driven by pharmaceutical companies desiring to give their product the best possible conditions to demonstrate efficacy, and by regulators wanting a 'rigorous', 'scientific' methodology. The ethical guidelines of the Council for International Organizations of Medical Sciences (CIOMS) (the primary international agency on biomedical sciences) express concern at the implications of the tendency to be selective in terms of participation in medical trials. It says, 'In the past, groups of persons were excluded from participation in research for what were then considered good reasons. As a consequence of such exclusions, information about the diagnosis, prevention and treatment of diseases in such groups of persons is limited. This has resulted in a serious class injustice' (CIOMS 2002). The attitude of national regulatory authorities is of particular interest as it would seem to go against this.

One interesting aspect of this discussion is the extent to which it echoes long standing concerns in the literature on technology transfer. For example, Chambers (1997: 70), commenting on the impact of the success of the Green Revolution in terms of aggregate crop outputs, notes that it 'reinforced beliefs that agricultural scientists knew what was good for farmers in all conditions … it was as though the research station … conditions were reproduced on farmers' fields. But this is not feasible for the complex, diverse and risk-prone agriculture of most resource-poor and rain-fed farmers'.

Table 3: Explanatory and pragmatic trials

	Explanatory Trials	Pragmatic Trials
Question	Efficacy: Does intervention change target outcome?	Effectiveness: Does intervention 'work' in normal practice?
Context	Tightly controlled, well-resourced.	Normal practice
Participants	Highly-selected, excluding those with conditions that might dilute effect and potential poor adherents.	Little or no selection beyond clinical indication of interest.
Intervention	Strictly enforced and adherence closely monitored.	Applied flexibly depending on specific patient.
Comparator	'Placebo' strictly enforced and adherence closely monitored.	'Normal care' applied flexibly depending on specific patient.
Outcomes	Often short-term surrogate and/or process measures.	Beneficial outcomes for specific patient
Relevance	Indirect: little attention given to relate to the decision-making processes of those using intervention in practice.	Direct: designed to meet the needs of those making decisions about treatment options in practice contexts.

Source: adapted from Zwarenstein et Treweek 2009.

Several recent articles highlight that RCTs are not immune from the risks of spurious findings – either positive or negative – from 'sub-group analysis' (the post-hoc examination of large-scale trial data to determine if results for the overall sample hold for particular groups within that sample). Randomised trials are such that repeated analysis by different sub-groups is very likely to eventually give rise to an interesting (statistically significant) finding, purely by chance. Wittes (2009) highlights these risks, citing the example of such an analysis of the value of aspirin in preventing fatal heart attacks which (specifically to illustrate the point) identified contrary effects for those born under the astrological signs of Gemini or Libra.

Aulakh and Anand (2007) discuss the need for caution, even when considering such 'obvious' candidate variables for sub-group analysis as age-group and gender. They cite another well-known randomised trial on the value of aspirin in preventing stroke. In this case a post-hoc sub-group analysis by gender was interpreted as indicating that beneficial effects were limited to men. This resulted in inappropriate advice being issued to clinicians in 1980 which was not corrected until 1998, following numerous studies that reported contrary findings. Wittes (2009: 913) spells out the dilemma generally researchers face: 'If reporting on subgroups is tempting but treacherous, failing to report on them seems unscientific and incurious.' She suggests a number of procedures, for example being extremely cautious when considering small sub-groups (which, for statistical reasons, are the most problematic), to reduce the risks of over-interpretation.

Theory based evaluation: realistic evaluation and theories of change

The European Commission Guide compares 'experimental' and 'realist' paradigms, possibly conflating what in practice are two distinct approaches to evaluation: theories of change (ToC) (Connell et al. 1995; Fulbright-Anderson et al. 1998) and realistic evaluation (RE) (Pawson and Tilley 1997). Both derive from the work of Weiss (1972) on theory based evaluation (TBE), which argued, primarily on pragmatic grounds, that 'social programs are based on explicit or implicit theories about how and why the program will work' (Weiss 1995: 66) and that an in-depth understanding and exposition of those theories was essential for designing rigorous monitoring and evaluation frameworks. The need for such an approach was based on a common perception that although a very large and increasing number of evaluations were being undertaken, there was very little evidence of their use. One reason for this impasse was that even where policymakers were open to the use of 'evidence' provided by traditional a-theoretical approaches, it was often inconclusive or contradictory, allowing selected evidence to be quoted to support predetermined positions.

Realistic evaluation

Realistic evaluation is a direct descendent of theory based evaluation; however, its popularity in the European context is partly explained by the extent to which it has been shaped by a European variant of the broader philosophical movement known as critical realism. Within this context, theory construction is seen by realist evaluators not simply as having methodological advantages but as a core requirement. For example, an evaluation of nursing practices states that: 'It is not enough to "prove" that a particular nursing intervention results in a positive outcome (e.g. improved satisfaction) for patients, … the reasons for these improvements also need to be understood' (Wilson and McCormack 2006). In their landmark book, Realistic Evaluation, Pawson and Tilley argue that the underlying problem with the 'experimental' approach is not the practical difficulties involved in its application to social programmes but more fundamentally 'its weakness as science' (1997: 30). They point out that the use of 'treatment versus control' trials is largely confined to the relative narrow, though clearly important, field of testing products or services for efficacy and effectiveness. In general, scientific experimentation involves (a) articulation of a plausible theory and (b) testing that theory under carefully controlled conditions to determine if predicted processes are initiated and outcomes observed.

Realistic evaluation essentially applies the same scientific methodology to assessing social programmes. However, this is an intrinsically much more complex task. A programme is targeted at diverse individuals, who may respond in very different ways depending both on their specific circumstances and on their perceptions of that programme. Those individuals will themselves play a major role in defining the contexts within which various causal mechanisms may be generated during implementation. These contexts will also be strongly influenced by the broader socio-economic system within which the programme is located. This implies that there is no possibility of manufacturing the 'carefully controlled conditions' described above. 'In laboratories scientists create artificial conditions in which those causal mechanisms which they conjecture to exist will be activated. In the natural world, potential causal mechanisms will only be activated if the conditions are right for them' (Tilley 2000: 5).

A programme evaluator therefore needs to simultaneously explore 'its underlying mechanisms and its contiguous contexts' (Pawson et al. 2005). For example, a child health promotion programme may aim to provide information, encourage trust in local services and empower mothers to take healthcare decisions. Programme implementation may trigger different processes depending on the characteristics of targeted individuals and households (age, socio-economic status, cultural beliefs, family circumstances, etc.), and various community and societal factors (availability of services, intra-household income distribution patterns, cultural norms, etc.). Across such varied contexts, the same programme components might in some cases result in increased knowledge and improved attitudes, and in others promote intra-household disagreements that impact adversely on children's health.

A central assumption of RE is that almost all social programmes are inherently complex and that evaluators have to accept and deal with that complexity. They do allow for rare exceptions to this rule. 'There are occasions (e.g. the occurrence of singular program theory, the opportunity for rigorous researcher control of the 'treatment', clear

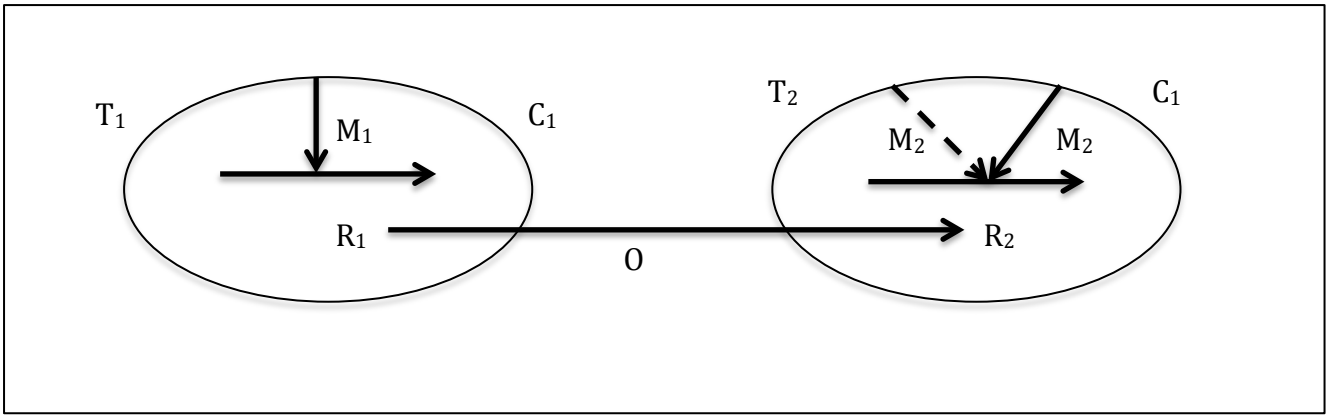
expectations about well-defined and immediate outcomes) when experimental principles can be followed usefully in evaluation research' (Pawson and Tilley 1998: 74). Experimental evaluation is considered to mask this underlying complexity by adopting a 'black box' approach (Victoria et al. 2004), where as long as 'outputs from' are consistently correlated with 'inputs to' there is no need to inspect the mechanisms behind those correlations. This means 'Everything needs to be apportioned as an "input" or "output"' (Pawson and Tilley 1997: 51), diverting attention away from the reality of the multiple processes involved in any intervention implementation.

A focus on observing the 'output response' to a given 'input signal' drives the evaluator to find ever more elaborate procedures to reduce the 'noise' in the system that will potentially mask that response. However, it may be precisely the characteristics of that noise that are of the greatest interest, both in terms of evaluative learning and in terms of evidence-based decision-making. For example, in the Mwanza exercise cited above, great care was taken to ensure that treatment and control groups were comparable, allowing a 'fair test' of the intervention. However, this methodological focus on determining whether or not the intervention had 'worked' encouraged the evaluators to bypass many equally or even more interesting and policy-relevant questions. For example: to what extent did outcomes vary across different types of individual and/or community? What were the key characteristics of the intervention or of those individuals and communities that lead to this variation?

Pawson and Tilley (1997) argue that the experimental approach also has a serious practical weakness as it fails to provide an explanation for contrary findings. They suggest this is a major reason for the limited use of evaluation results, even though so many are undertaken and there is such emphasis on the need for evidenced-based policy. The key barrier is the fact that evaluation findings in most fields are highly inconsistent – some showing that an intervention 'works' and an equal number reporting that it does not. Policymakers then either seize those findings which justify existing preferences or argue that the evidence is inconclusive and that it is only prudent to delay any policy changes until the position is clarified. By developing explicit and detailed theories to explain how intervention components caused a measured change, persuasive evidence with the potential to advance policy debates can be accumulated.

Within RE such theories are termed 'mechanisms' (M). They provide explanations for observed 'regularities' (R) within a given 'context' (C). These regularities may relate to relevant indicators (e.g. child malnutrition rates), empirical associations between two or more variables (e.g. educational attainment and poverty status), or observed patterns of behaviour (e.g. non-use of public health facilities by minority communities). Interventions are then seen as introducing and maintaining new mechanisms which induce desirable changes in those regularities. Such changes are designated 'outcomes' (O). Figure 1 shows a successful intervention in which the problematic regularity R1 in context C1, resulting from mechanisms M1, is transformed into the improved regularity R2 when the more powerful mechanism M2 is introduced. Note the assumption that the overall context remains (more or less) constant. The evaluator determines the extent to which this assumption was plausible in any given intervention.

Figure 1: The elements of a successful intervention



Source: Adapted from Pawson and Tilley 1997: 74, Figure 3.8 Basic ingredients of successful programmed social change.

The guiding 'rules' for a realistic evaluation, specified by Pawson and Tilley (1997) relate to attitudes and processes rather than methods:

1. Reject evaluation as a search for 'constant conjunction' between intervention and outcome. Commitment to theory-based evaluation and 'generative' causality (i.e. outcomes) are generated by intervention mechanisms interacting with individual and community responses.
2. Accept that many of the important factors in determining intervention success or failure are generated by a range of social processes at different levels and may not be immediately observable. For example, the responses of poor individuals resident in a rural community may be strongly influenced by extended family and other social networks extending far beyond the boundaries of that community.
3. Focus on contexts. An intervention mechanism's success will depend on the context within which it is introduced. Evaluators identify what works 'for whom and in what circumstances'.
4. Focus on mechanisms. Problems that an intervention addresses must be understood in terms of existing causal mechanisms that give rise to those problems. Also, how new mechanisms introduced interact with existing mechanisms, and the extent to which they have countered them has to be understood.
5. Outcomes need to be analysed rather than simply measured. The aim is to explain outcomes in terms of specific mechanisms acting in a given context.
6. Learning – developing transferable and cumulative knowledge – can best be achieved by focusing on CMO (Context + Mechanism = Outcome) configurations. An evaluation commences with an assumed CMO configuration which is refined by the evaluation process.
7. Evaluators should adopt a teacher-learner relationship with those initiating, implementing and participating in an intervention. It should be assumed that each of these stakeholders has considerable but limited understanding of the intervention.
8. Accept that the context into which an intervention is introduced is typically evolving and open to external factors.

Theories of change

Theories of change evaluators would agree with those adopting the RE paradigm on at least two key issues (Blamey and MacKenzie 2007). First is the crucial need for in-depth understanding of the given context (socio-economic, demographic, political, environmental, organisational, etc.) in which an intervention is introduced. Some elements of this context may be under the control, for example those implementing the intervention may obtain government assurances about not planning major policy changes that would disrupt intervention activities. However, most may not be under control, and evaluators have to allow for the possibility of substantive changes in the context. Second is the need to move away from the notion that an intervention can be viewed as a unified 'treatment' whose 'effect' can be ascertained by aggregation over highly heterogeneous groups of individuals. The expectation should be that different subgroups of the targeted population will present a range of responses to individual components of the intervention, given the specific contexts within which they encounter those components. Most, but not all proponents (Bezzi 2006) also agree on a third point, the rejection of 'constructivist' or 'fourth-generation' approaches to evaluation (Lay and Papadopoulos 2007). These are seen as proposing interesting methods which might be used to improve internal validity, for example by exploring the perspectives of different stakeholders, but failing to address external validity, counterfactuals or generalisation to wider populations (Connell and Kubisch 1998; Pawson and Tilley 1997).

Attempts to determine differences between RE and ToC are complicated by their adherence to exclusive terminologies. Blamey and MacKenzie (2007: 445) point out that the failure to agree terms 'which are sometimes used interchangeably and can often be counter-intuitive', has given rise to substantial confusion in the theory based evaluation literature. They identify two distinct categories of 'theory' which have been variously defined. The first relates to what Weiss called 'implementation theory', which addresses the processes whereby inputs are converted to outputs: 'what is required to translate objectives into ongoing service delivery and programme operation' (Weiss 1995: 58). The second, which Weiss calls 'programme theory' (later referred to as 'causal theory' [Weiss 1998]), addresses the mechanisms from the interactions between intervention components and targeted population, and the likely outcomes. For example, the implementation theory for a new micro-credit scheme would have to provide a basis for estimating staffing levels to undertake the number of face-to-face transactions involved to be able to meet the demand for services. The programme theory would focus on the likely response levels of target groups to the services and the mechanisms (e.g. perceived opportunities for higher income or increased security), which would motivate them become or stay as members.

Blamey and MacKenzie suggest that, while RE and ToC approaches acknowledge the importance of the two interacting components in principle, for many ToC practitioners 'uncovering programme theory is perhaps more aspirational than practical' (2007: 445). Thus, evaluations described as ToC have tended to emphasise implementation theory (also noted and regretted by Weiss [1997]). Indeed, in a number of practical applications labelled as ToC the model often appears to map the implementation process by essentially elaborating the logical framework approach (LFA), with the standard links between inputs, activities, outputs, outcomes and impacts extended to show more complex causal pathways (e.g. Geddes 2006). There may be attempts to justify some of

the key causal links but these rarely identify the specific indicators that will determine if a given link has been triggered, an essential requirement of a ToC evaluation. For example, although we can assume that financial incentives will make health providers more willing to work in remote areas, a ToC implementation theory would have to specify the level of incentive required to achieve this. Its programme theory would have to postulate the likely responses of different types of providers.

One reason for this focus on implementation theory can be deduced from Box 1, which outlines the steps in an evaluation as seen from the theories of change (ToC) and realistic evaluation (RE) perspectives. The ToC evaluator places great emphasis on participatory articulation of the underlying intervention model. This involves seeking facilitated agreements between all stakeholders on a set of activities which, if properly implemented, will lead to a desired set of outcomes. There is greater attention on '*types of activities, timescales and anticipated outcomes or thresholds of change*'.

Box 1: Comparative guidelines for theories of change and realistic evaluations

Theories of change evaluation (ToC)	Realistic evaluation (RE)
1. Stakeholders agree the long-term vision 2. Stakeholders consider the necessary outcomes required by the end of the initiative. 3. Stakeholders articulate the types of outputs and short-term outcomes required to achieve the specified targets. 4. Implementers consider the most appropriate activities or interventions required to bring about the required outputs and short-term outcomes. 5. Stakeholders consider the resources that can realistically be provided. 6. Evaluators interrogate the implied theory of change to ensure that the under lying logic is acceptable to stakeholders either because of its existing evidence base or because it seems likely to be true in a normative sense. 7. Evaluators interrogate the implementation theory to ensure that timescales, financial resources and capacities are appropriate.	1. Evaluators, through dialogue with stakeholders, attempt to understand the nature of the intervention: aims; target population; contexts; and prevailing theories about why initiative will work for some people in some contexts. 2. Evaluators map out a series of potential 'mini theories' that relate the various contexts of an intervention to the multiple mechanisms by which it might operate to produce different outcomes. 3. Evaluators undertake an 'outcome inquiry' in relation to these mini theories. This involves building up a quantitative and qualitative picture of the programme in action, partly through dialogue with stakeholders. It might, for example, address how patients with different types of chronic illness respond to different kinds of treatment delivered in a variety of ways. 4. Evaluators, through an exploration of how context, mechanism and outcome (CMO) configurations play out within a programme, refine and develop tentative theories of what works for whom in what circumstances.

Source: adapted from Blamey and MacKenzie, 2007: 443 – 444.

Substantial resources are allocated to this process because 'it is these stakeholders who best understand the intervention and it is they who will, at a later stage, require to be convinced that the outcomes that are measured are attributable to the detail of the intervention theory they approved'. They will also need to 'be involved in key decisions about which … elements of the theory become the foci for the actual evaluation activity' (Blamey and MacKenzie, 2007: 442). When these tasks are done, the evaluator can explore the underlying rationale and plausibility of the proposed model. This would include programme theory relating, for example, to the changing knowledge, attitudes and behaviours of individuals or groups participating in the intervention. With time and resources constraints, the scope for undertaking the in-depth analysis required to complete this, often more problematic second phase, is limited.

Realistic evaluators are primarily concerned with programme theory. The focus on exploring CMO scenarios that generate responses in intervention participants is so that it provides a basis for programme theories (to explain why individuals or groups respond in a particular way) (Stame 2004). Such explanations are unlikely to be available at the intervention level. Beneficiary responses can be mapped by specific intervention components only if they come into contact with them. In the micro-credit example above, a realistic evaluator might consider a preliminary interview with an individual an important component for the scheme's operation. Depending on the context (such as age, gender, ethnicity, socio-economic status and perceptions) a range of responses could be

generated ranging from trust and optimism to hostility and despair. This could in turn generate a range of behaviours that could determine the success or failure of the scheme for specific sections of the population.

Realistic evaluation accepts that a clear understanding of the basic implementation theory is also required, but only as a necessary step to identify and analyse CMOs. As with the ToC approach, a range of stakeholder perceptions and attitudes are sought. But here this is to inform the evaluator about the CMO configurations to be prioritised and appropriate methods to be adopted.

The focus on CMOs marks a clear distinction between RE and ToC in terms of the approach to attribution and causality. ToC evaluators are open to the possibility that even a complex intervention may be described as having 'worked'. This will typically be at the level of 'plausibility' (Table 1). If the various stakeholders agree on the expected (preferably quantitative) outcomes for each stage, and those targets are attained, then attribution of outcomes to intervention would be acceptable. However, following this logic, ToC evaluators would have no objection to those who sought 'higher-levels' of proof, for example through randomised controlled trials or quasi-experimental trials (Weitzman et al. 2002). An RE evaluator however, would regard this as a foolish and costly pursuit of the unattainable.

For example, a ToC evaluator would require the implementation theory underlying a randomised trial of a community-based health insurance scheme to be mapped out and agreed by all stakeholders. A RE evaluator would consider collecting information on non-treatment districts a waste of scarce resources. RE would involve a detailed investigation, for example, a screening interview to identify people for subsidised care. This would establish how the interview would influence outcomes for various sub-groups (gender, age, ethnicity, etc.). A programme theory based on previous experience and observing large number of interviews would then help in re-designing interviews, either for this intervention or the next.

1.3 Theory into practice?

Three widely held theoretical positions can help clarify the underlying perspectives of the above debate:

1. Experimental: RCTs provide the only scientific approach to evaluation. It may sometimes not be possible to undertake such trials, in which case we should strive to approximate the RCT benchmark as closely as possible by very careful construction of a counterfactual. However, we accept that this is very much a second-best option.
2. Theories of Change: RCTs would be the best option in an ideal world but it will usually be impossible to employ them in practice. With or without RCTs, it is essential that we focus not simply on whether an intervention succeeded or failed but why. By devoting sufficient resources to developing a shared understanding on how an intervention works, we can design monitoring systems that will allow us to evaluate the extent to which observed outcomes can be plausibly attributed to the intervention. Where feasible, the use of a RCT or well constructed counterfactual can provide valuable supporting evidence.
3. Realistic Evaluation: Interventions are multi-faceted and highly malleable, with various components adapting to local contexts. Participants in the implementation process, including beneficiaries and managers, will have a diverse range of characteristics, perceptions and attitudes that shape responses to these components. Placebo effects will be large and uncontrollable. External factors will also give rise to unforeseen effects that vary over the intervention period. Given this reality, it is irrational to seek evidence that a particular type of intervention works. RCT or 'quasi-experimental' designs are a waste of time and resources in terms of systematic learning. It is only possible to identify the most interesting specific intervention components and explore how their performance in relation to specific groups or individuals. This allows programme theories to be constructed that genuinely advance our knowledge, and help in modifying current interventions or design new ones.

One underlying distinction between these three positions relates to the different weights that they explicitly or implicitly attach to the various evaluation objectives. Realistic evaluation tends to focus almost exclusively on the need for systematic learning, rarely addressing issues of accountability. The extent to which an intervention has succeeded or failed is of limited interest, as it cannot provide reliable insights for future interventions of a similar type. The experimental approach is very much concerned with these issues.[2]

[2] The first principle of International Initiative for Impact Evaluation (3ie) (www.3ieimpact.org) states that: '3ie supports impact evaluations that adhere to agreed-upon methodological standards for addressing the "attribution challenge" – e.g. establishing cause and effect between programmatic activities and specified outcomes.'

Genuine RCTs can determine an intervention's impact. They provide assurance against the effects of confounding factors and selection bias that encourage the use of probability (as against purposive sampling) in statistical surveys. Clinical RCT trials of healthcare treatments are one of the powerful scientific tools available, contradicting long-held beliefs based on observational and epidemiological studies. In one meta-analysis, Ioannidis (2005) found that 5 out of every 6 findings of such studies could not be replicated through other methods.

One common criticism of RCTs, made by ToC and RE practitioners, is that the emphasis placed on determining the success or failure of an intervention can distract attention from understanding the underlying mechanisms contributing to that result. However, from a practical perspective the demand that we must know 'how' an intervention works seems excessive.

Historically, the vast majority of healthcare treatments, disinfectants, fertilisers, pesticides, etc. were adopted on the basis that they worked, long before there was any understanding of the processes involved. Even today, new drugs are routinely bought into service before their precise action on the body has been determined because they have been assessed (using RCTs) as safe and effective. The theory becomes important when findings from repeated studies are inconsistent. RCTs in varied contexts showing income gains for farmers given access to a new technology, would be a rational basis to encourage that technology, even if how those gains had been achieved is not entirely established.

There are two serious concerns with the experimental approach. First, the methodological status given to RCTs may lead to an uncritical assessment as to what is required to meet the strict assumptions underlying such a claim. For example, it is often not possible to specify experimental and control populations at the individual level. Most exercises involve cluster RCTs, which randomise at the level of geographical areas such as villages or districts. Resource constraints often lead to the use of a small number of large clusters. Statistically, this could result in very large theoretical sampling errors that undermine the robustness of findings. More problematic are 'quasi-experimental' designs, which adopt methods such as propensity score matching (Ravallion 2002) and assume that absolves them from critical analysis and interpretation of findings.

Note that one particular characteristic of 'gold standard' clinical trials, double blinding (where neither researchers nor participants know the membership of treatment and control groups), seems to have been conveniently ignored by those advocating the experimental approach in other areas. Placebo effects cannot be disregarded simply because double blinding is infeasible. The placebo effect has been shown to have both complex and substantial influences on treatment outcomes (Blasi et al. 2001). A review of complementary medicine trials (Ernst et al. 2008) indicates that non-blinded trials were much more likely to provide evidence of successful outcomes. The impact of an agricultural project also could similarly be influenced by the responses based on whether participants are in the treatment or control group, irrespective of the substance of that project. Similarly, enthusiastic project managers could go out of their way to advance the situation of the experimental group, even where this involves going beyond the project terms of reference.

The second concern in relation to the experimental approach is that 'the implementation is the intervention'. As noted above,, the implementation of even an apparently simple technical intervention in agriculture involves a complex social project. Despite detailed and precise project designs, the interaction of its components with the diverse perceptions and attitudes of the target population and other stakeholders generate a unique set of 'contexts and mechanisms'. RE advocates would argue that this implies that the project should also be seen as a unique experiment that can never be replicated. For example, new crop marketing arrangements may be seen as a welcome opportunity by some and as a threat by others. The balance between these groups within a community, the strength of feeling in each group and the extent to which the community has mechanisms for resolving such conflicts have a decisive effect on project outcomes. Such factors will have been largely determined by the specific history of that community and will thus vary substantially across communities. A or quasi-experimental evaluation might demonstrate that the new arrangements lead to aggregate benefits, though an analysis of 'winners' and 'losers' would be needed. However, as the results are specific to the studied communities, there will be doubts as to the likely outcome if the exercise were repeated with a different target population. A series of such exercises might result in wholly inconsistent findings, with the new arrangements giving rise to substantial benefits or losses, depending on the population under study.

Theory of change and realistic evaluation proponents would argue that such conflicting results are far more typical than the long series of positive (or negative) findings discussed above. Both would identify the atheoretical nature of the experimental approach as the underlying problem. With conflicting outcomes and no intervention theory to guide us, we reach an impasse. A ToC evaluator's response would be to develop a model that will allow us to determine why the intervention works in some cases and not others. The first step would be to understand how the intervention was intended to function (the implementation theory) and then to map this against actual

performance, identifying divergences and bottlenecks in the causal chain (from inputs to outcomes). From an RE perspective, having developed a basic understanding of the implementation theory, the aim would be to identify the key mechanisms that determine outcomes for specific population groups in specific contexts – the programme theory. For example, we might explore the process whereby purchasing prices are determined and how in practice this process is applied to and perceived by different sub-groups within the community, for example richer and poorer or men and women. This knowledge can then be used to design new interventions that are more appropriate for specific populations and contexts.

If RCTs, or well-designed quasi-experimental studies, provide the most persuasive evidence as to the impact of a specific intervention and ToC or RE offer alternative approaches to systematic learning, it makes sense to adopt a 'combined methods' approach so both accountability and learning objectives can be satisfied. As indicated above, those following the ToC paradigm, unlike their RE counterparts, have no theoretical objection to the use of RCTs or quasi-experimental designs. There would be resource implications. The use of treatment and control groups is only useful to the extent that reliable comparative data on changes over time are collected, analysed and interpreted for both. It has been suggested above that the emphasis ToC places on seeking stakeholder agreement on detailed implementation theory tends to constrain attempts to develop programme theory. Both attempting to meet the requirements of an experimental design approach alongside work to develop both implementation and programme theory runs the risk that inadequate resources will be allocated to at least some of these activities, and stakeholders may be confused. The act of data collection also generates its own politics.

1.4 An alternative approach?

While evaluators debate these theoretical and philosophical distinctions, evaluation as such is at risk of being sidelined, especially where the accountability and implementation agendas claim priority. Nielsen and Ejler (2008: 172) argue that, 'from the perspective of many public managers and fieldworkers, evaluation reports often produce a body of knowledge that appears too late and is too long … to be useful as a management tool. Yet at the same time they demand timely information on the results of the programme delivered. It is a problem that some consider a profound challenge to the very role of evaluation in the knowledge society'. They note that the changing demands of policymakers, combined with technological advances which allow routine computer-based monitoring, have tended to displace traditional evaluation practices in favour of what has come to be known as 'performance based monitoring' (PBM) or 'results based monitoring and evaluation' (RBME).

PBM is intended to generate the information required to both demonstrate and enhance 'value for money'. The objective is not attribution but 'ascertaining that the politically intended social value has been created' (p. 177). The emphasis is on accountability – providing evidence that allocated resources are correlated with quantifiable benefits – and performance. 'It is the linking of implementation progress (performance) with progress in achieving the desired objectives or goals (results) of government policies and programs that makes results-based M&E most useful as a tool for public management' (Rist 2006 p 4-5). From the point of view of those implementing an intervention, demonstrating these links will usually be sufficient to gain the approval of both their peers and the population at large.

Table 3 compares attributes of PBM and formal evaluation. The appeal of performance based monitoring (PBM) is clearly reflected in the current focus on 'outcome-orientation' in the design of monitoring systems for development projects and programmes, and the specific reference in the Paris Declaration to the importance of 'results-oriented reporting and monitoring frameworks' (OECD DAC 2005: 8). This approach is also seen in the monitoring of Poverty Reduction Strategies (discussed in Section 2). In spite of its attraction to many donors, the approach has significant risks. It can be potentially difficult to design and implement effective PBM systems even where adequate resources are available. A recent Global Fund evaluation found that its version – performance based funding – had 'evolved in practice into a complex system that focuses primarily on short-term metrics addressing mainly project inputs and outputs as opposed to development outcomes and impacts. Further, while the system generates extensive data, it often fails to provide the key elements of information required to inform judgments on effectiveness.' (Sherry et al. 2009: 30). (Annex 1 provides an extended discussion of issues relating to Global Fund monitoring and evaluation).

Table 4: Comparison of evaluation and performance based monitoring

Phase	Item	Evaluation	PBM
Design	Purpose	Negotiated up front	Evolves over time
	Scope	Specific issues	Broader focus
	Budget	Separate budget item	Integral to programme
	Frequency	Episodic	Ongoing
	Timing	During/after programme	Through programme
	Units of measurement	Customised quantitative and qualitative indicators	Quantitative indicators reproduced over time
	Type of indicator	Input, output, outcome, impact	Input, output, outcome
Obtaining data	Production	One time	Routine processes
	Tools	Desk research, interviews, surveys, information systems	Surveys, information systems
Analysis	Means	Triangulation of multiple sources	Limited sources
	Tools	Contribution analyses, time series, regression analyses, experimental designs	Contribution analyses, time series
	Attribution	Attribution often a key objective	Attribution assumed
Evaluative judgement	Tools	Benchmarking, cost-effectiveness, cost-efficiency, multi-criteria analyses, expert panel and many other	Predominantly bench-marking, cost-effectiveness, cost-efficiency
	Performance standard	Descriptive or prescriptive	Descriptive
	Assessor	External or internal programme evaluators	Internal programme managers
	Format	Evaluation report	Tabular reporting with short explanatory text
Utilisation	Organisational learning	Low to medium	High
	Budgeting cycle	Occasionally	Mostly
	Users	Few	All levels of programme organisation
	Tactical decision-making	Low	High
	Strategic decision-making	High	Low to medium

Source: adapted from Nielsen and Ejler 2008: 175.

An alternative approach to either of either of the three approaches discussed above, and one probably also acceptable to advocates of RE , would be to address the accountability issue using performance based monitoring. This would specifically be intended to generate the information required to both demonstrate and enhance 'value for money'. Combining the micro-level in-depth learning approach of RE with the 'is the intervention achieving its targets and allocating resources efficiently?' objectives of PBM, is an interesting possibility. The absence of a control group may be seen as a serious objection by some. However, many policy makers may be perfectly happy with the 'adequacy' level of inference discussed above, which requires only that convincing evidence be provided of the achievement of intended outcomes.

2 The practice of M&E in Poverty Reduction Strategies (PRS)

The World Bank, IMF and many other international donors along with 67 countries prepared Poverty Reduction Strategy Papers (PRSPs). The rationale of this was based on what became known as 'process conditionality'. This involved 'opening up discussion among stakeholders within developing countries about ways and means of addressing poverty reduction goals', with an emphasis on: country ownership; comprehensive scope – both across sectors and in terms of potentially available resources; and 'performance-based' or 'outcome-oriented' in terms of resource allocation (Booth and Lucas 2004). The underlying logic of the PRSPs was that substantial funding would be available for countries that were willing to propose:

(a) a comprehensive strategy for poverty reduction that donors found acceptable and
(b) a plausible monitoring strategy that would reassure that that strategy was being effectively implemented and producing the intended outcomes.

PRSP M&E systems are of interest for two main reasons. First, they emphasise the essentially political nature of M&E. Multiple stakeholders with different perspectives and interests agree to a joint statement on what constitutes accountability and evidence of performance. Many of those stakeholders are aware that there is little possibility that existing or proposed monitoring systems will be able to reliably deliver most of the indicators which form the basis of that agreement. This is usually not critical because no one wishes to see the PRS 'fail'. There is an implicit understanding that donors will rarely withhold funding unless the recipient government 'behaves badly', for example failing to introduce promised legislative reforms. An evaluation process which is often presented as an exercise in results based monitoring typically becomes a series of negotiated compromises, based on whatever limited evidence is available, around issues that will need to be addressed in the next round of funding.

The willingness of senior policymakers to enter into agreements based on often overly-optimistic returns from M&E, has frequently had practical consequences. Confronted by demands from their superiors for data from barely functional routine information systems or infrequent and relatively small-scale sample surveys, those responsible for delivering the PRSP indicators (often with the assistance of externally funded consultants) were driven to adopt a range of innovative procedures to generate plausible alternatives. Some of these are considered below.

2.1 The Poverty Reduction Strategy context

Two basic principles of the PRSP process are of particular relevance. The initial guidelines (World Bank/IMF 1999) required that the strategies should be:

- ***Country-driven and owned*** – meaning that the process will be led by the government but include the broad-based participation of civil society in the adoption and monitoring of the poverty reduction strategy.
- ***Results-oriented*** – the strategy needs to set medium and long-term goals for poverty reduction, including key outcome and intermediate indicators to ensure that policies are well designed, effectively implemented and carefully monitored.

In the PRSP conceptual framework 'poverty reduction' was conceived not simply in terms of increasing incomes or expenditures but more broadly, as 'increased well-being'. While the macro-economic chapter of PRSP generally reflects commitment to a growth strategy focused on reducing income poverty, considerable emphasis was placed on other aspects of poverty reduction via the provision of access to services such as health, education, water and sanitation. They also frequently addressed issues of empowerment or security through the promotion of policies on decentralisation, the role of civil society organisations, or the provision of safety nets. This reflected the role many international agencies allocated to PRSPs in attaining the Millennium Development Goals (MDGs); the PRSP was described as 'the 'national roadmap' for reaching long-term MDG targets' (UNDP/World Bank 2002; see also Renard 2006). PRSP outcome indicators were specifically intended to reflect this multidimensional view of poverty.

One reason for this approach was the complex relationships between indicators relating to various dimensions of poverty. 'There is little correlation, for example, between poverty reduction and changes in under-five mortality, and none at all between non-income MDGs, such as primary education and under-five mortality. Surprisingly, there is a strong correlation between poverty reduction and changes in underweight, but virtually none between poverty reduction and under-nourishment' (Bourguignon et al. 2008: 4). Also, as Klasen (2005) points out, neither income nor expenditure measures can address concerns relating to intra-household poverty. They can only identify poor people on the basis of their membership of poor households. On the other hand many non-income indicators – malnutrition, literacy, enrolment, etc. – are based on the characteristics of individuals. It is important

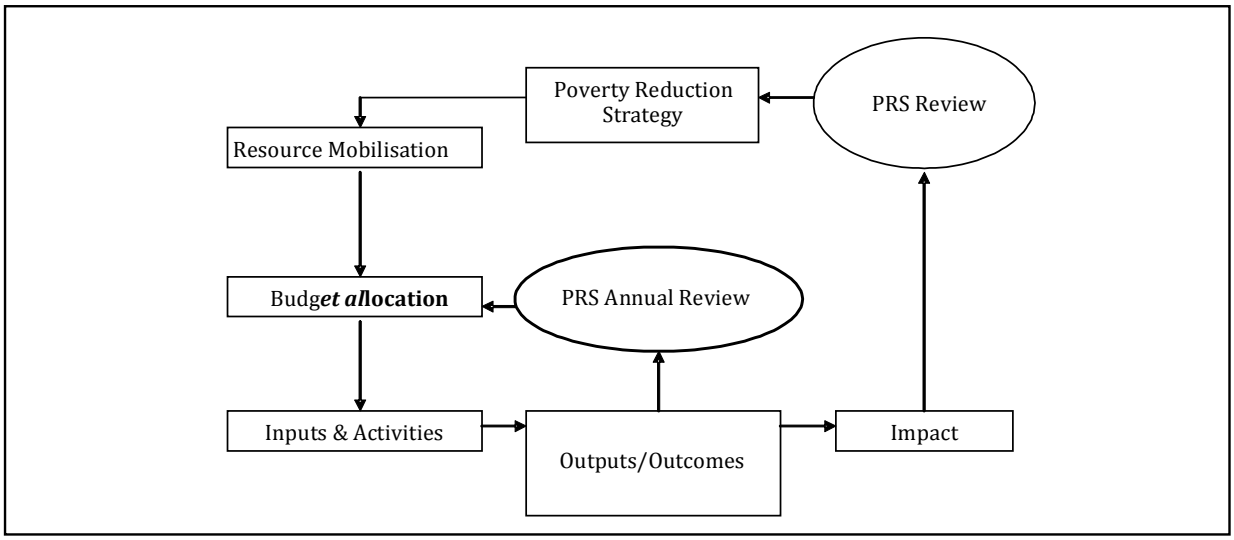
to check if reduced household income-poverty results in increased welfare for all household members, irrespective of their age, gender, status, etc. Non-income poverty indicators can provide at least some valuable insights into this area. In practice, however, PRSP M&E frameworks rarely follow through on this logic. Thin et al. (2001: 8), in an early review said, 'Despite evidence of multi-dimensional poverty concepts in all documents, income poverty is everywhere taken as the basic definition of who is poor. Thus we find phrases like 'the poor are more likely to be less educated' and 'malnutrition increases with poverty', but not 'the poor are more likely to have less income'. This omission paradoxically shows that lack of education and lack of adequate nutrition are seen as less basic to the definition of poverty than lack of income.'

Some countries are now in the third round of these multi-year, wide-ranging, strategic plans that are subject to internal and external review annually. The World Bank/IMF's acceptance of a PRSP requires a detailed discussion of the intended monitoring and evaluation procedures that will provide the evidence to these reviews. The 'PRSP Sourcebook' (Prennushi et al. 2002), developed by the World Bank and IMF to guide the implementation of the initiative, specifically highlights the need for a plan which involves non-government actors in these activities: 'nongovernmental actors – research institutions, civil society organizations, special-interest and advocacy groups, and others - have an important role to play in the design of the monitoring and evaluation system, in actually carrying out monitoring and evaluation activities, and in using the results'. Several countries went further to prepare detailed M&E plans and identify existing institutions (or establish new ones) that would take responsibility for their implementation. Many international and bilateral donor agencies have provided support to specific activities under these M&E plans, often providing substantial funding, though with mixed results (Holvoet and Renard 2007).

M&E in the PRSP Sourcebook follows a traditional view, using a revised version of the project logical framework approach (LFA) that is extended to cover programmes and policy changes. The basic conceptual framework (

Figure 2) illustrates the annual and overall planning cycle of PRS programmes and projects. Note that evaluation takes place at two levels: an annual review tied to the budget cycle, and a 'final' review after five years, typically leading to the production of a new PRS proposal.

Figure 2: PRS programme/project cycle



Source: adapted from PRS Monitoring Plan for Sierra Leone, January 2005.

The related M&E framework focuses on identifying:

(a) indicators to the various stages of the above cycle (Table 5);
(b) methods used to measure these indicators;
(c) institutional arrangements to collect, compile and disseminate the required information.

Table 5: Monitoring and evaluation indicators

Type of indicator	Used to Measure
Final outcome / Impact	Improvements in well being of the population
Intermediate outcome	Changes in income levels, consumption of quality services, better social and governance conditions and other factors directly affecting well being
Output	The immediate results of activities implemented under poverty-reduction policies – e.g. school buildings and trained teachers
Process	The institutions, procedures and operational mechanisms adopted to implement interventions
Input	The delivery of funding and other necessary resources and conditions for agreed activities to the responsible institutions

Source: adapted from 'Uganda Poverty Monitoring and Evaluation Strategy', Table 2.1, April 2002.

Three broad approaches to the PRS M&E framework development that the funding agencies adopted include:

1. An 'arms-length' strategy, restricting indicators to financial inputs, outcomes and impact. The aim here, in line with the Paris Declaration (OECD DAC 2005), was to entrust the PRS implementation to relevant government agencies and collaborating actors. As long as progress on a small number of key outcome/impact indicators could be demonstrated (Adam et al. 2004), development partners should avoid all involvement in operational and managerial issues.
2. The above can be a high risk option for both sides. National governments may agree targets for medium term outcomes that prove unattainable if, for example, there is a downturn in the international economy. Donor agency officials may feel obliged not to intervene even if they become convinced that these targets have no chance of being met. One alternative is to engage much more with process and output indicators that allow routine monitoring of the performance of implementing agencies. However, given the limited capacity for M&E in most countries, this typically results (contrary to the principles set out above) in a major shift towards 'donor-driven' systems. This is especially so as any additional demands for information will almost certainly not be met without additional donor support and funding.
3. Also a subset of process indicators (milestones) is important in practice. These are agreed actions to be taken by a given date, for example, the establishment of a central PRSP monitoring agency, the implementation of competitive tendering arrangements in the health sector, or curriculum reform in primary schools. These are often linked to the 'good governance' agenda and to the government's 'commitment' to the PRS process. This is implemented using a performance assessment framework (PAFs) or Policy Matrix (Booth et al. 2005), which is reviewed quarterly or twice a year. As these reviews are directly linked to fund disbursement, they are higher priority than the annual PRSP monitoring reports. Two main criticisms of this approach are that performance in the PAF may take precedence over performance in terms of delivering real benefits to the population. Secondly, it embeds those funding the PRSP firmly into the national policy process, potentially leading them to be insufficiently self-critical if agreed policies fail to deliver.

None of these strategies has proved wholly satisfactory. A recent review suggests that 'Donors seem to be somehow caught in a chicken–egg dilemma. As long as a minimum institutional capacity in terms of design, implementation and evaluation apparatus is not installed and functioning, the move towards new aid instruments which shift responsibilities to recipients may well be resisted by the more sceptical donors, and even those that go along with the new approach may still chose to duplicate the recipient country's fledgling systems with their own, with related demands on recipient systems that go a long way to undermine the whole approach' (Holvoet and Renard 2007: 2). These authors also endorse a long standing concern about the 'missing-middle' in PRSP monitoring systems: the insufficient attention given to monitoring the assumed causal chain(s) linking inputs and outcomes, often because they were poorly understood (Booth and Lucas 2002). They suggest that the results-based management approach, promoted by many donor agencies (Booth, Christiansen and de Renzio 2005; White 2005b), prioritised inputs and final outcomes, while neglecting the operational pathways that determine if those outcomes will be achieved.

Many remain sceptical about the prospects for establishing more effective, country-owned PRSP M&E systems. However, there is a general acceptance that it has at least resulted in an increased interest in data production, analysis and interpretation, increased allocation of additional resources to these activities, and an increased

openness to innovative methods. There is likewise an increased willingness to address issues such as 'trust' and 'empowerment' which are resistant to simple quantification. Various innovative qualitative methods have been devised or adopted and PRSP M&E has been a major area of activity for those involved in the 'qual/quant' movement (Kanbur 2003), which has been exploring effective ways to combine quantitative and qualitative methodologies.

Much of the innovation has been driven by necessity. PRSP M&E design and implementation were often challenging in developing countries where, in stark contrast to the European context, reliable data on almost any topic was unavailable. Earlier 'poverty monitoring' exercises with which many PRSP countries were familiar had essentially involved compiling available information using a similar approach as that of producing national statistics reports. PRSP's demands were radically different. First, whereas a poverty status report combined the most recent estimates for its various sources, all the information in a PRSP M&E report had to relate to the specific time period over which assessment was being made. Second, data had to reflect progress in terms of both short-run poverty reduction and PRS implementation. These two factors taken together required moving away from a focus only on final outcome indicators to including many more input, process and intermediate/output indicators. Third, those responsible for M&E had to convince others (including a possibly sceptical donor audience), that the data was reliable and sufficiently sensitive to assess the changes over one year. Fourth, presenting findings, for example at an annual review, had potentially serious consequences. At the very least, progress would be applauded and lack of progress questioned. At worst, failure to meet agreed targets could place future funding at risk.

Initially, support for large-scale surveys targeting outcome indicators, especially those relating to the MDGs, was emphasised. Recently the emphasis has shifted to the need for less resource-intensive methods which can deliver findings relatively quickly. Considerable attention is focused on approaches which engage and empower communities and beneficiaries. Because the basic PRSP principles stress country ownership and widespread civil society involvement, participatory monitoring and evaluation (PM&E) is being emphasised, especially in relation to monitoring and evaluating PRSP implementation (Schnell and Forster 2003). This was an essential mechanism whereby civil society organisations, on behalf of the population at large, could hold government accountable, though this premise has often proved to be overly-optimistic (Parkinson 2009). Some of the best known examples, which aim at bringing together local communities with policymakers and service providers to assess and where necessary modify PRS interventions include:

- citizen report cards for education and health services;
- traditional participatory poverty assessments (PPAs);
- qualitative impact monitoring (QIM); and
- community score card methodologies.

Many of these methods can be interpreted as belated attempts to address the 'missing middle' problem described above. Booth and Lucas (2004) argue that the root cause of this failure lies in simplistic assumptions as to the linear and self-contained nature of the intervention process: policy – implementation – effects – outcomes – impacts. In reality, intervention design and implementation are inextricably combined, with crucial decisions sometimes taken by minor actors far removed from the original policymakers; unpredicted effects are the rule rather than exception; and 'other things' are almost never equal, with specific contexts (location, population, institution) playing a major role in determining eventual impacts. A predetermined framework of indicators and targets remains an essential starting point but may be of limited value in the absence of an in-depth appreciation of the complexities of the implementation process. This implies a need for much greater flexibility in the types of information gathered and the methods by which it is generated, analysed and interpreted. New forms of M&E are required which recognise the significance of monitoring implementation of programmes, and feeding back information which support researchers, policy makers and other stakeholders respond to what is happening' (Mosse et al. 1998).

2.2 A review of selected methods

Many countries slated to participate in the PRSP initiative had previously received support to undertake routine 'poverty monitoring' exercises, based on data from household expenditure surveys or integrated household surveys which included an expenditure module. Familiarity with the poverty monitoring methodology (i.e. the tendency to associate 'outcome-orientation' with a focus on outcome indicators and a continuing fascination with income-poverty) led to an assumption that similar surveys would play a key role in PRSP M&E. However, it was quickly apparent that the relatively high cost per household (which restricted it to a small sample size) and the long turnaround times for such surveys limited their potential use. The annual PRSP review process ideally required the routine delivery of key indicators on a similar time scale. Monitoring PRSP initiatives implementation

by local government agencies similarly demanded a high level of disaggregation and samples large enough to allow reliable estimation at this level. Most countries continue to undertake large scale surveys over the overall PRS period, with routine M&E often reliant on regular surveys that are specifically intended to track policy/outcome linkages on an annual basis. Some of the alternative methods that were the result of innovations to address these issues are presented in this section.

Core welfare indicators questionnaire (CWIQ)
The core welfare indicators questionnaire (World Bank 1999) was developed by the World Bank, UNICEF and the UNDP. It is an off-the-shelf household survey package designed to provide a means by which leading intermediate/output PRS indicators can be estimated annually using a relative simple survey instrument on a sufficiently large sample of households (typically 20,000 to 100,000 households) to allow for disaggregation on at least a regional, gender and age basis. It includes questions allowing the estimation of a set of common standardised indicators on well-being and access to, utilisation of and satisfaction with basic services. The questionnaires are machine readable, and analysis fully automated allowing results to be generated within weeks of the fieldwork.

Public expenditure tracking surveys (PETS)
Public expenditure tracking surveys (Lindelow et al. 2001) are targeted at institutions and were designed primarily to track budgeted expenditures down to the units providing services (schools, clinics, etc.). The primary aim is to estimate the proportion of such expenditures reaching each unit and how these eventually translate into service provision. Because they involve detailed work at facility level, a range of data on inputs, outputs and quality are collected which are directly relevant to PRS requirements. They are sometimes linked to quantitative service delivery surveys (QSDS). These are targeted at service delivery units such as health facilities and schools, and were initially developed as a research activity within the World Bank (Dehn et al. 2001). They are intended to be complementary to PETS, focusing on quantitative performance data derived mainly from the routine data system in each unit.

Citizen report card surveys
Citizen report card surveys (ADB 2007) were introduced as a way of gathering and disseminating systematic feedback from users on the quality and performance of government service providers, to highlight gaps and bottle-necks in service delivery and to improve accountability. The specific methodology predates the PRSP initiative and was first introduced by the Public Affairs Centre in Bangalore, India in 1993 (Catholic Relief Services 2003). The underlying concept is much older, dating back to at least to Florence Nightingale in 1863: 'I am fain to sum up with an urgent appeal for adopting this or some uniform system of publishing the statistical records of hospitals. If they could be obtained . . . they would show subscribers how their money was being spent, what amount of good was really being done with it, or whether the money was doing mischief rather than good.' (quoted in McNamara 2006: 101). The survey follows a standard consumer market research methodology, asking a limited number of simple factual or perceptual questions which may relate to an individual provider or more generally to the services offered in a given geographical area. It also requests basic respondent demographic data. The findings, presented in the form of a school report card, consist of aggregate scores (means, percentages, etc.) relating to the perceived performance (availability, access, quality, etc.) of different services, aspects of a single service, or even aspects of a single provider. In some cases the user survey data may be combined with that from a parallel provider assessment which can augment the scorecard with simple factual data relating to quality standards (Table 6).

Table 6: Selected quality indicators used in Uganda and Bangalore report card survey

Technical quality – structural	Availability of water, Uganda Waste disposal mechanisms, Uganda Drug stock management procedures, Uganda Client registries, Uganda Containers for needle disposal, Uganda Frequency of visits for which doctor was present, Bangalore
Technical quality – process	Adherence to guidelines for monitoring growth of children, Uganda Management of malaria cases, Uganda
Interpersonal quality – outcome	Waiting time, Uganda and Bangalore Patient privacy, Uganda Patient's level of satisfaction with doctor, Bangalore

Source: adapted from McNamara 2006.

A recent development of this approach is the 'balanced scorecard', which has been found useful in countries where routine data systems are highly unreliable (Peters et al. 2007). This draws on the work of Kaplan and Norton (1992), who advocated the need for assessing enterprises using a multi-dimensional framework that considered the relationship between current operational activities and longer term objectives. Given that service provision, for example in the health sector, is equally complex and has multiple components and objectives, the aim of the balanced score card in this area is to encourage policymakers and managers to resist the temptation of pursuing some activities and neglecting others by providing a transparent and simple overview of progress across a range of performance indicators. These indicators are groups under various 'domain'. For example:

1. patient and community outcomes,
2. staff outcomes,
3. capacity for service provision,
4. service provision outcomes,
5. financial system outcomes, and
6. overall vision outcomes.

In a participatory process, stakeholders jointly select indicators within each of these domains using the criteria of face validity and importance, while also taking into account reliability, comprehensiveness, outlying values and variation. For each indicator, upper and lower benchmarks are set to indicate achievable targets, for example by considering the best and worst performing regions or providers. Estimated scores are derived from various surveys and expert assessments. They are displayed on a 'performance dashboard' (Table 7) which highlights regions or providers which fall outside the lower and upper benchmarks. These display charts may reflect current performance or changes in performance over time. A variant on the general approach involves estimation of the performance scores by the providers themselves as an internal exercise designed to highlight areas for collaborative improvement (HPI 2008).

Participatory poverty assessments

Participatory poverty assessments (Norton 2001; Robb 2002) were initially seen as a natural counterpart to expenditure surveys within the PRSP monitoring framework. In a number of countries the sampling frames for the two approaches were deliberately aligned to allow comparative exercises. However, triangulation of findings between surveys and PPAs proved problematic, with many arguing that the scope for direct comparability was limited. PPAs might be helpful in terms of raising survey design issues but the different concepts of poverty that they embodied meant that they should rather be seen as complementary approaches (Appleton and Booth 2001). In some countries this led to the use of PPA exercises to address issues arising from household surveys, with findings feeding back into the design and analysis of those surveys. A natural extension of this approach was the applying PPAs to explore specific PRSP implementation issues, with a particular focus on the investigation of delays and bottlenecks (Ssewakiryanga 2005).

Table 7: Balanced score card indicators: health services in Afghanistan

Domain A: patients and community	District 1	District 2	District 3	Etc.
1 Overall patient satisfaction %				
2 Patient perception-of-quality index %				
3 Reported community activities %				
Domain B: staff				
4 Health-worker satisfaction index %				
5 Salary payments current %				
Domain C: capacity for service provision				
6 Equipment functionality index %				
7 Drug availability index %				
8 Family planning availability index %				
9 Laboratory functionality index (hospitals and CHCs) %				
10 Staffing index–meeting minimum staff guidelines %				
11 Provider knowledge score %				
12 Staff received training in last year %				
13 HMIS use index %				
14 Clinical guidelines index %				
15 Infrastructure index %				
16 Patient record index %				
17 Facilities having tuberculosis register %				
Domain D: service provision				
18 Patient history and physical examination index %				
19 Patient counselling index %				
20 Proper sharps disposal				
21 New outpatient visit per month (BHC >750 visits) %				
22 Time spent with patient (>9 minutes) %				
23 BPHS facilities providing antenatal care %				
24 Delivery care according to BPHS %				
Domain E: financial systems				
25 Facilities with user fee guidelines %				
26 Facilities with exemptions for poor patients %				
Domain F: overall vision				
27 Females as % of new out-patients %				
28 Outpatient-visit concentration index CI (–1 to 1)				
29 Patient-satisfaction concentration index CI (–1 to 1)				
Composite scores				
1 Percent of upper benchmarks achieved %				
2 Percent of lower benchmarks achieved %				

Source: Adapted from Peters et al. 2007: 148

Qualitative impact monitoring

A closely related approach to issue-oriented participatory M&E has been supported by the Deutsche Gesellschaft für Technische Zusammenarbeit organization (GTZ) in a number of countries. Qualitative impact monitoring or consultative impact monitoring (CoIMPact) methodologies (Schnell and Forster 2003; GTZ undated) are based on a 'community immersion' paradigm and attempt to engage a team of relevant stakeholders, including national and local government staff, local institutions and community representatives in an extended intervention-focused research exercise. This consists of four phases: collaborative problem identification and study design; fieldwork involving a wide range of participatory methods involving a stay of at least 5 days in each community; quantitative and qualitative analysis of findings; and a variety of dissemination activities tailored to meet the requirements of all

stakeholder groups. One advantage of the approach suggested by Asche (2003) is that it allows 'politically challenging questions' to be raised in a context that encourages direct engagement between policymakers, those responsible for implementing policy and intended beneficiaries.

Community score card process
A participatory counterpart to the PETS and citizen report card methodologies has been used in a number of PRSP countries, mainly with World Bank support. The community score card process (World Bank 2005) was designed to allow communities to assess the resources available to local service providers (compared to government specified entitlements), and the performance of those providers. Entitlements and actual resources are first assessed collaboratively by facilitators and providers. In a health facility, this process would include undertaking inventories of equipment and drugs and examination of available financial and service provision records. The findings are then presented to community representatives for clarification and debate. The Input Tracking Matrix, displaying expected against actual resources is developed at this meeting. The second stage involves collaborative identification of a set of key performance indicators. This is usually undertaken by focus groups of service users. A final agreed list of 5-8 indicators is then compiled from the proposals of these groups. Many of these indicators will be similar to those used in the citizen report card approach and cover both perceived and factual information (e.g. opening hours of the clinic) and user satisfaction. The focus groups then allocate scores on these indicators, using some form of voting mechanism. In parallel, a focus group of provider representatives score their own performance on the same set of indicators.

The primary objective of these procedures is not to generate performance data, but to open up discussion between providers and users about key performance concerns and possible ways to remedy these. The input tracking exercise is intended to allow some measure of agreement as to whether externally imposed resource constraints can be blamed for poor performance. If so, then there may be scope for collaborative action between providers and users to demand action from local officials. On the other hand, if performance is below what is expected given available resources, there may be possibilities for negotiation between providers and users to seek solutions that are acceptable to both sides.

Poverty and social impact analysis (PSIA)
The PRSP initiative has also been closely associated with recent work on poverty and social impact analysis, which assesses the impact of complex policy interventions on the well-being of different stakeholder groups. It focuses on the multiple potential pathways by which these impacts might be transmitted. Numerous case studies of interventions have been published over recent years, many about the agriculture sector, by different national and international agencies. An additional output are three substantial 'toolkits': 'Evaluating the Poverty and Distributional Impact of Economic Policies' (Pereira da Silva et al. 2003); 'Tools for Institutional, Political and Social Analysis of Policy Reform' (Holland 2007); and 'The Impact of Macro Economic Policies on Poverty and Income Distribution' (Pereira da Silva et al. 2008). These constitute a valuable reference on a multitude of traditional and innovatory, quantitative and qualitative methods.

PSIA was intended to provide the core research component of the PRSP process, attempting to understand the detailed implications of specific PRSP interventions, especially in terms of distributional outcomes and the impact on poor and vulnerable people. At one stage the World Bank indicated that every substantial policy initiative should be subject to an ex-ante PSIA though this ambition has proved wildly optimistic (Holvoet and Renard 2007; Bird et al. 2005). It draws on a wide variety of tools and methods from a range of disciplines. The full range is detailed in the three toolkit volumes. In terms of its theoretical approach to evaluation, PSIA can be very much seen as following a theory based evaluation rubric, providing a detailed description of the 'links in the chain' between intervention and impact on poverty, clearly setting out all the assumptions. The World Bank Users Guide (World Bank 2003) advocates a public statement of initial assumptions relating to:

- the channels that will deliver (positive or negative) intervention impacts,
- how different institutions and agents are expected to respond,
- exogenous conditions required for the intervention to be successful,
- and any exogenous risk factors.

As in the discussion of theories of change evaluation, this is intended to allow an informed debate between stakeholders, which can in turn be the basis for a negotiated agreement leading to any modifications or refinement of the underlying model. It should also provide the basis for designing effective and timely monitoring procedures and feedback mechanisms that help adjust implementation.

The theory based evaluation influence can also be seen in the emphasis placed on context in a PSIA. 'The design of reforms is based on underlying assumptions about the context and the behavioral response of key institutional

and human actor'" (World Bank 2003: 15). The initial phase of analysis provides for a detailed examination of this context, with a particular focus on institutional and stakeholder analysis (World Bank 2008). The former addresses not only the specific institutions involved in the initiative but also the broader political, social and economic environment within which they function. The latter maps out a diverse range groups the intervention can affect, or who can exert positive or negative influence over its implementation. The analysis focuses on the characteristics of each these stakeholders, assessment of their attitudes and opinions, identification of incentives which make them to support or oppose the intervention, and the nature and extent of their capacity to influence the implementation process or outcomes.

The central PSIA analytical framework can be seen as an extension of the poverty impact assessment (PIA) matrix developed by the Asian Development Bank (ADB) for project appraisals (ADB 2002). This provided a short summary analysis of the logic underlying the ADB project on income-poverty reduction. It considered four 'impact channels' (now termed 'transmission channels') that reflected the various income-flows that the intervention might affect: labour earnings of household members in the workforce; prices in markets where they sold or purchased goods; access to and return on non-labour household assets; and net receipts of public and private transfers. The rows of the PIA matrix identified one of these impact channels. The columns described the nature of the impact: direct, indirect (i.e. via another impact channel), or macro (for example, via general price inflation) and its time frame: short, medium or long term. Given the PRSP stance on the multi-dimensional poverty concept, the concept of assets was extended to include 'livelihood assets' (physical, natural, human, social and financial) (DFID 1999) and a fifth channel capturing 'access' to goods and services, for example, schools and health facilities, was added. More recently, the framework was further extended to allow for impacts which came about via changes in the 'rules' under which stakeholders operated. This is now generally called the 'authority' channel. Table 8 provides a recent version of the PIA matrix (OECD DAC 2007).

Table 8. Poverty impact assessment matrix: Transmission channels and outcomes for target population

Transmission Channels & Details		Transmission Channel Used	Output/Outcome/Impact by Transmission Channel Categories			Information Sources (S)
		Details & Risks (T)	Short Term (+/-)	Medium Term (+/-)	Details & Risks (D)	
Prices	Production					
	Consumption					
	Wages					
Employment (includes self-employment)	Public formal					
	Private formal					
	Informal					
Transfers	Taxes					
	Public welfare/subsidy					
	Private remittances					
Access	Public services					
	Other					
Authority	Formal organisations					
	Informal relations					
Assets (change in returns and/or in levels)	Physical					
	Natural					
	Human					
	Social					
	Financial					

KEY Table 1:	Strength/direction impact	+ +	+		-	- -
		very positive	positive	not significant	negative	very negative

Further discussion of PSIA and an example of its application are provided as Annex 2.

2.3 Key lessons from the PRSP experience

Underlying much of the discussion around PRSP M&E systems are two key facts. First, in most of the countries involved there were severe capacity constraints across the range of skills required. Second, these constraints were not sufficiently recognised or acted upon. From a purely technical perspective, the poor quality of the available (or potentially available) data should have implied that a conservative attitude was adopted as to what could plausibly be achieved, and there was willingness to make hard choices in term of prioritising monitoring activities. In practice, it more often resulted in a version of the 'don't ask, don't tell' strategy, with all sides agreeing to gloss over many of the evident shortcomings of the M&E system. Given the emphasis in the Paris Declaration on 'outcome-orientation', 'results based monitoring' and national ownership of M&E systems, it was very difficult for either side to suggest that there was simply insufficient capacity to deliver on the promised performance indicators or that the original monitoring framework had been wildly over-ambitious. As a result, the annual reviews typically adopted the less confrontational alternative and assembled quantitative and qualitative information, making a best-estimates about the progress on agreed targets and attempting to identify concerns and challenges.

The awareness that 'second-best' options would have to play a major role in the M&E process provided a fertile environment for developing innovative approaches to data collection. This was reinforced by a realisation that 'outcome-orientation' need not imply a single-minded focus on outcome indicators. Where there was a widely accepted theory of change, intermediate output/outcome monitoring (in addition to input monitoring and tracking of resource flows) could be seen as a rational and cost-effective way to assess progress. It also opened up possibilities for rapid feedback on implementation, which in contrast to retrospective assessments based on long run final outcome measures, could potentially influence future behaviours and procedures, as suggested by the discussion on performance based monitoring above.

While the innovative methods discussed had considerable potential, in practice, they were often poorly implemented and under-resourced. Rapid surveys intended to deliver reliable basic data on a few key indicators had to include several and often more problematic, questions because this seemed an easy and inexpensive option. Participatory exercises were undertaken by inexperienced junior researchers (or even statistical office field staff) following a one week training course in 'participatory methods'. Overall there was often a failure to appreciate the three-way trade-off between data complexity, data quality and resources.

3 Conclusions

3.1 Is the 'bad' press for agriculture M&E justified?

ALINe's assessment of the existing state of M&E in agricultural interventions (Lindstrom 2009; Haddad et al. 2010) may seem familiar to many working in other areas: a 'compliance culture', expressed in a preoccupation with accountability to donors rather than to intended beneficiaries and other stakeholders; failure to fully integrate M&E into intervention planning and implementation processes; limited capacity among those responsible for M&E; limited understanding of its potential value among other staff (M&E primarily seen as an additional burden); and insufficient resources to deliver findings of an appropriate quality. These are common complaints, with underlying causes linked to deeply entrenched attitudes that either attach limited importance to accountability and transparency or are reluctant to allocate the often substantial resources required to achieve them. To some extent they probably also reflect a failure on the part of the evaluation community, either in terms of providing convincing evidence of the value of their activities or in finding effective methods to promote them.

However, research by ALINe suggests that agricultural interventions may have intrinsic characteristics that make particular demands on M&E (see Haddad et al. 2010). These include:

- a lack of clarity as to primary objectives – projects typically have multiple objectives entailing complex tradeoffs;
- long 'causal chains', in terms of both number of links and overall project duration (Millstone et al. 2010; and
- sensitivity to uncertainties imposed by climate and other natural phenomena, accentuating the potential disconnect between individual incentives and programme impacts (Sabates-Wheeler et al. 2010).

The overall implication seems to be that the theory based chains of causality from agricultural development projects to hunger and poverty impacts are complex highly non-linear and subject to substantially higher levels of risk compared to health and other social sectors. The resulting difficulty in specifying the 'implementation theory' (Weiss, 1995) of such interventions seriously impedes the design of an appropriate M&E system. In the absence of a realistic model of the process by which an agricultural intervention is intended to translate inputs into clearly identified outcomes, it is very difficult to know how to monitor or evaluate performance.
Is the position substantially worse than in other sectors? It is perhaps natural to believe that one's own area of study poses very special methodological problems. The author works mainly in the health field and has always assumed that those working on the education sector have a far easier time. In terms of the above comparisons, leaving aside the highly contentious issue of DALYs (Anand and Hanson 1998), it is true that health interventions do have the overall purpose of improving health status. However, the health sector has struggled with the theoretical and practical difficulties inherent in the measurement of this elusive concept (e.g. Mortimer and Segal, 2008). Such difficulties have led the great majority of health projects to adopt various mortality-based impact indicators and a range of proxy outcome indicators such as access, utilisation and quality of services, all of which raise serious definitional and measurement issues, and are not affected by health 'inputs' alone. From a health perspective, the concern with multiple objectives in agriculture would seem more than offset by the availability of reasonably well defined and potentially measurable variables to assess some of those objectives. Many health sector evaluators would look enviously on indicators such as crop output, yield per hectare, market value of production, nutritional status and even household income per capita.

The argument relating to the relative complexity and length of causal chains in agriculture is more compelling. Again making comparison with the health sector, there are a wide variety of well understood basic health interventions that are generally regarded as both effective and inexpensive. The primary concern is that health services in many developing countries seem unable to deliver them, particularly for poor people. The implementation theory for these interventions is reasonably well-defined and plausible causal chains specified. Yet progress in many areas, for example reducing maternal mortality, has been painfully slow. One key lesson would seem to be that even apparently simple, evidence-based, medical interventions typically entail complex social interventions that require concerted and innovative efforts to understand, engage and incentivise a diversity of stakeholders. This has lead to a focus by a number of researchers on what is described above as intervention programme theory – attempting to understand how specific types of individuals respond to different aspects of an intervention within a specific context.

This can be seen as a shift away from previous mainstream evaluation work in the health sector, which has broadly adopted the experimental approach. For instance, an editorial in The Lancet (2004) which applauded the increased attention given to RCTs for programme evaluation by the World Bank was titled 'The World Bank is finally embracing science'. There is also solid support for the experimental paradigm in both the European and

US policy evaluation communities. The 'Coalition for Evidence-Based Policy' (www.coalition4evidence.org), a not-for-profit organisation based in the United States that includes many leading academics, was established specifically to address what they saw as the serious problem that 'social programs are often implemented with little regard to rigorous evidence, costing billions of dollars yet failing to address critical needs of our society'. In seeking such evidence they 'limit this discussion to well-designed randomized controlled trials based on persuasive evidence that they are superior to other study designs in measuring an intervention's true effect'.

On the other hand, the recent World Bank publication discussed above, which reviews a number of recent health interventions, argues that the 'applicability of RCTs to "treatments" that involve complex strategies, including most approaches to strengthening health services, is limited' (Peters et al. 2009: 11). On a similar theme, the author was recently involved in the evaluation of a ten year health-sector reform programme in China which covered 71 counties and included a diverse package of specific components which varied substantially both between counties and over time. The Chinese evaluation team devoted substantial resources to the post-hoc construction of a counterfactual, using propensity score matching (Ravallion 2002) to identify comparison counties. The decision to undertake this activity was made primarily to conform with what were seen as donor preoccupations: 'In the results-based climate of today, impact evaluations focus on outcomes. In any study they conduct, evaluators should be concerned with the requirement to take account of the counterfactual' (White 2005a: 10). Many of those involved (and probably many of the external evaluators) regarded this exercise as a purely formal activity, with little real value in terms of assessing the merits of the programme. They would almost certainly have been much more inclined to support another comment by White: 'Constructing a picture of how the intervention has played out on the ground, which nearly always requires data from the treatment group alone and not from a comparison group, is essential to a good impact evaluation' (undated: 10).

3.2 What can we learn from the PRS process?

While there is little discussion of evaluation theory as such in the PRSP M&E literature, it is evident that the issues encountered and approaches adopted are very similar. With the exception of RCTs, which are absent from the PRSP literature, all of the theoretical areas discussed in Section 1 of this paper seem relevant. The 'outcome orientation' stance of the PRSP Sourcebook and the widespread implementation of citizen report cards are clearly influenced by the performance-based monitoring approach and its U.S. equivalent. The ever more complex logical frameworks and flow diagrams included in PRSPs and the resources allocated in many countries to stakeholder involvement in the design of the various poverty reduction strategy components indicate a least a leaning toward the theories of change position. Many of the qualitative and participatory exercises at community level, especially those involving detailed investigations of bottlenecks in the implementation process, can be seen as realistic evaluation type exercises.

These exercises were often introduced because there were no other sources of information on key areas of the PRS. As discussed above, those responsible for producing the PRSP and negotiating funding with the donors often agreed to implement an M&E strategy that was far beyond what their existing data collection and reporting systems could deliver. Statistics and information departments that had struggled to provide aggregate outcome indicators with a three or four year time lag were suddenly confronted by a need to provide current, detailed process data that would be examined annually by an expert team of donor-funded consultants. Given that conducting traditional large-scale surveys or upgrading existing routine information systems were infeasible options, given the time pressure and resource constraints, donors and country PRS monitoring teams collaborated to develop alternative and cost-effective procedures.

These procedures included a range of both quantitative and qualitative methods[3]. For example, the core welfare indicator questionnaire, and similar 'rapid' surveys, covered a strictly limited set of key output indicators, used a short machine-readable questionnaire and aimed for a two-four week turnaround time. The 'bottleneck' participatory poverty assessments and similar exercises involved a series of community based activities focusing on a key issue which was seen as threatening the progress of a given intervention. In The Gambia, an exercise of this type overturned a generally accepted assumption that rural Muslim parents would not send their girl children to school for fear of them coming into contact with male students and indicated that a small compensation for the labour time foregone would greatly advance the female recruitment process. The overall approach might be characterised in term of what Chambers (1986) has called 'optimal ignorance' or 'proportionate accuracy' (i.e. seek the minimum affordable information required for good decision-making) and is relevant to the debate around

[3] Set out in the toolkits described above: 'Evaluating the Poverty and Distributional Impact of Economic Policies' (Pereira da Silva et al. 2003); 'Tools for Institutional, Political and Social Analysis of Policy Reform' (Holland 2007); and 'The Impact of Macro Economic Policies on Poverty and Income Distribution' (Pereira da Silva et al. 2008)

the use of quantitative methods in agriculture, often to the detriment of qualitative methods. Hence potential effects that cannot be easily measured in this way will tend to be missed, and when not missed the qualitative methods used will not be state of the art and even when they are the audience may not be entirely receptive.

Perhaps one of the most interesting lessons of the PRSP experience is the very clear demonstration that M&E is not a technical activity but 'a fundamentally political one with technical dimensions'(Booth and Lucas 2004). It is political in the sense that, as stressed in both the theories of change and realistic evaluation literatures, it involves a series of complex negotiations between the various stakeholders. In the PRSP example, policymakers in many countries welcomed the proposal that they should take 'ownership' of the M&E process but were constrained by human resource capacity. Donors were initially reluctant to become involved in the M&E process, but felt they might not have appropriate information to demonstrate accountability to their own constituencies. These issues led to the ultimate reliance on process and milestone indicators. Many of the previous technical and capacity issues that plagued the PRSP process early on have been addressed; however, the purpose of the PRSP M&E – what should it be trying to achieve and for whom – remains a contentious issue..

For an M&E system to play an effective role in this process, mechanisms which clearly link the information flows which it generates to such incentives, have to be established. Some types of information are much more suitable for this task than others. Many of the 'gold-standard' poverty outcome indicators, for example poverty prevalence, mortality rates, illiteracy rates, etc. are in practice, of little value because it is very difficult to attribute responsibility for changes to a specific agency or individual, and they are usually available on a time-scale which has no relevance for current policy. The author worked in one country where a simple error resulted in the publication of a survey report showing a threefold increase in poverty prevalence in two years. The only comments on this publication related to 'unfavourable conditions in the global economy' over this period. At the other end of the scale, in some countries the annual publication of national school examination results can be a major political event for district level authorities, with poor performance sometimes resulting in sharp declines in the popularity of local politicians and demands by parents for the replacement of key officials.

Such examples imply that one key aspect in the design of M&E systems should be a detailed consideration of the extent to which changes in the selected indicators have clear and specific consequences for identified agencies and/or individuals. However, two caveats are required. First, if information is to have serious consequences, rather than to be seen as merely 'interesting', then it is essential that it be reliable. Many agencies at both national and international levels have repeatedly proved hopelessly optimistic as to the possibilities for generating the timely, high quality data that such an approach would demand. Second, such a system would create both 'losers' and 'winners'. It could only be implemented where senior politicians and officials were prepared to deal with the potentially serious antagonism which this would inevitably generate. It is not clear what incentives would persuade either national governments or donor agencies to go very far down this path.

3.3 Some final thoughts

In a discussion of the evaluation 'paradigm wars', Pawson and Tilley (1998: 73) suggest that these have resulted in 'a musty stalemate over first-principles, which has made no difference to the overall balance of evaluation activities'. Much more productive, they argue, would be a debate around the routine decisions involved in specific instances: 'much is to be learned by comparing alternative research designs for a particular evaluation. Similarities and differences can be highlighted, and strengths and weaknesses of differing strategies identified.' Key to such debates is the fact that the different stakeholders' priorities – justification of resource allocation decisions, accountability for resource use, improved implementation management, learning, etc. – will vary considerably, and that different evaluation designs are more suited to some objectives than others. Given that an evaluation involves some element of trade-offs between objectives, there is a need in terms of evaluation design for transparency in setting out those objectives and a realistic assessment of the extent to which each can plausibly be achieved.

One relevant observation from the realistic evaluation paradigm is that what might be seen as simple interventions from the supply-side (e.g. providing improved access to a new vaccine or crop variety) will almost invariably be seen as complex interventions from the demand-side. In any intervention, target population groups and individuals will have diverse interests and perceptions that result in very different responses to what are assumed to be identical stimuli. This is not simply a question of considering the probable responses of, for example, men and women, young and old, rich and poor, urban and rural, etc. Variation within such groups may be much the same or even greater than variation between them, based on individual personalities, histories and circumstances. Interventions will almost always trigger a raft of unpredictable, context-specific mechanisms and outcomes. This

implies that design and implementation are inextricably combined and that each intervention is to a substantial extent unique. Those designing evaluations have to deal with that reality.

M&E is essentially a political activity with technical aspects. The PRSP experience suggests that both donors, who made unrealistic demands, and country policymakers, who made unrealistic promises, were to some extent trapped by the prevailing rhetoric that demanded both rigorous performance measurement and the primacy of national information systems. There was rarely any attempt at systematic assessment of the practicability of the proposed PRSP M&E plans, which were only seldom addressed by the World Bank officials overseeing the PRSP negotiations (Booth and Lucas 2004). This game playing around M&E was probably to the detriment of both sides. The objectives of donors were met – but only on paper. The review teams were rarely presented with reliable evidence on the agreed targets and had to manage with whatever was available. On the other hand, national information systems remained weak and under-resourced. The opportunities for enhancement and innovation in those systems, which could, at least in principle, have been used to drive an incentive-based implementation management process – with resource flows linked to routinely monitored performance – was missed.

4 Annexes

4.1 Annex 1: The Global Fund and Performance-Based Funding

The Global Fund to fight AIDS, Tuberculosis and Malaria (GFATM) was launched in 2002 with the aim of targeting both public and private resources to address what were regarded as the three most important diseases afflicting developing countries. It shares two unusual attributes with the Global Alliance for Vaccines and Immunisation (GAVI) which was established two years earlier. First, it has no permanent staff in the countries to which is provides funding. Design and implementation of programmes is managed by country partners, which a 'local fund agent' (LFA) is contracted to undertake auditing and performance assessment. Second, and most important in the present context, it has adopted a system of performance-based funding under which resources are transferred on the basis of demonstrated progress against agreed coverage targets (Feachem and Sabot 2006). The first section of the GFATM M&E manual states that: 'The principle that M&E is mostly concerned with is Managing for Results, which, for the Global Fund, translates into Performance-Based Funding (PBF). Effective PBF means that: Grants are invested where the greatest impact on HIV, TB or malaria can be achieved; Grant recipients have strong incentives to focus on results and timely implementation; and Programs and the Global Fund can identify what works in a particular program for early replication, systems strengthening and scale-up (learning from results)' (Global Fund 2008: 3).

A funding proposal is submitted to the Global Fund Secretariat by a Country Coordinating Mechanism (CCM), composed of government, NGOs and civil-society organisations. If it is accepted, following the recommendation of an independent technical review committee, an agreement is negotiated with a 'principal recipient' selected by the CCM. Each such agreement will include a Performance Framework, a contractually binding document which defines an agreed list of indicators and targets that will be used to assess performance and thus determine access to additional funding (Global Fund 2007). The Secretariat will also require the submission of a detailed M&E framework. This will cover a range of indicators in addition to those in the Performance Framework and provide baseline values, targets, sources, collection procedures, frequency of collection and responsible agencies. It will also describe how the information compiled will be verified and quality assured, preferably using an approach developed by the Global Fund (2008), and how it will be disseminated to stakeholders and the general public. Finally, it will contain an agreed Action Plan, specifying M&E activities and outputs, with details of budgets, agencies and timing. It is recommended that 5-10% of total funding should be allocated to M&E.

The implementation of the PBF mechanism is described as: Raise it; Spend it; Prove it; Raise it (Global Fund 2008: 4). Funds will typically be allocated to cover the first two years of a programme. Towards the end of the second year an evaluation will be undertaken covering: progress in terms of the targets specified in the agreement; LFA assessments relating to procurement, M&E (with a particular emphasis on data quality) and implementation progress; and contextual issues which may have affected progress, including natural disasters, political unrest or economic crisis (Radelet and Siddiqi 2007). The Secretariat will use this information to assign an overall evaluation score: A, B1, B2 or C as shown in table 1. On the basis of these scores a decision will be taken to: continue funding; continue funding with some additional conditions attached; continue funding only if targets and budgets are substantially changed; or discontinue funding.

Table 1: Global Fund programme evaluation scores

A: Meeting or exceeding expectations	**B1: Adequate**	**B2: Inadequate but potential shown**	**C: Unacceptable**
Number of people reached with services			
Targets met or exceeding 80%	Significant improvements made (50–80%)	Some improvements (30–50%)	Marginal or no improvements (<30%)
If the programme has achieved significant improvements in terms of numbers of persons reached, the Global Fund does not consider lower-level indicators for the Phase 2 decision			
Number of service centres established or strengthened			
		Significant improvements (>30%)	Marginal or no improvements (<30%)
Number of people trained to deliver services			
		Significant improvements (>30%)	Marginal or no improvements (<30%)

A number of authors have examined the approach adopted by the Global Fund. Radlet and Cairnes (2005) in a report for DFID, applaud the development of a detailed 'Monitoring and Evaluation Toolkit' (The Global Fund 2006) which provides standardised indicators for AIDS, Tuberculosis and Malaria, and note substantial improvements over time. However, they list a number of concerns including: ability of the technical review panel to assess health systems; lack of detailed country contextual information; the burden imposed by quarterly reporting requirements; limited capacity of LFAs to implement data quality verification; and difficulty of attributing results in a multi-donor context. Murray et al. (2004: 1099) suggest that 'the pressure at the national level to provide biased data will intensify' as agencies link disbursements to achievements and point to issues raised in a data audit commissioned by the GAVI (LATH Consortium 2001). Brugha et al. (2002: 438), are also concerned that the 'evidence suggests that performance-linked rewards will induce exaggerated reports of achievements at different levels of the system'.

Most recently, the five year evaluation of the Global Fund (Sherry et al. 2009) had identified a number of 'considerable limitations' associated with the PBF system. While acknowledging that it has encouraged a results oriented approach and an 'internal culture of accountability', found that it had 'evolved into a complex and burdensome system that has thus far focused more on project inputs and outputs than on development outcomes, departing from the vision of an outcome-based model. Most importantly, there remain inadequate information system and monitoring and evaluation capacities in countries critically limiting the feasibility of the performance-based funding approach' (2009: 29).

The underlying issue relates to the need for a more realistic appreciation of the very low capacity of many national health information systems. Quality assurance guidelines which appear rational and appropriate where the need is to 'fine tune' information systems that are not functioning as well as they might, can appear fanciful when applied to systems that are barely functional. Efforts to improve the situation often 'had the unintended consequence of making the system more confusing at the level of implementation' (p. 30). They were also of limited practical value given that the most serious quality concerns related to baseline data. The evaluation suggests that where there is no practical short term means of raising the quality of data to the level required for the LFA and technical review panel to operate under existing guidelines, the realistic option must be to revise those guidelines. They suggest that it should be possible to introduce 'more differentiated approaches to quality assurance that are capable of improving performance and accountability monitoring within existing capacity constraints in countries' (p. 29).

Many of the other main issues raised by the evaluation team can be seen as arising from the well-recognised tendency of performance-based contracts to encourage funded agencies to focus exclusively on contracted targets even if this involves diversion of resources from activities of equal concern in terms of the overall health system. Of particular concern was the admission by implementers in the majority of countries that they had sometimes 'sacrificed quality of implementation in order to achieve a quantitative numerical PBF output target' (p.31). Sub-contracting by the principle recipient was seen as resulting in a 'fragmentation along disease lines', with responsible agencies implementing disease-specific surveys, clinic reporting forms and even information technology systems that result not only in inefficient allocation of resources and additions burdens for health providers and managers but often multiple incompatible data sources. Other donors are seen as seen as complicit in this failure to adopt the appropriate systemic approach, 'mainly investing in devising separate monitoring and evaluation strategies, seeking consensus around core indicators, and stepping up reporting requirements for countries' (p.31).

4.2 Annex 2: Poverty and Social Impact Analysis

The focus on transmission channels in PSIA can be seen as one way to address the complexities of TBE described in the main text. Rather than track the effects of the multitude of individual mechanisms generated by the various components of an intervention, PSIA explores the net effect of those components in terms of each of the transmission channels, asking about the implied changes in prices, employment, etc. that are triggered by intervention inputs. Typically the number of such primary transmission channels will be limited for any given intervention. The analysis can then repeat the above process, moving on to consider to what extent these changes trigger second round effects via the same set of transmission channels. For example, an irrigation scheme may directly increase farmers' income via the employment channel, as production rises, and the value of their land via the assets channel. The increased production may trigger second round changes via the prices channel (e.g. reduced production and increased agricultural input prices) and the assets channel (e.g. increased borrowing using land as collateral and/or increased investment in production equipment). The aim will be to track this 'results chain', focusing on the most important effects as the number of branches inevitably increases with each round, and noting the timescale over which each effect operates. The analyst will also have to take account of the fact that an intervention is not a closed system and that there will therefore be a increasing need to include external factors in the evaluation as time passes.

The definition of transmission channels is clearly somewhat arbitrary. The aim is to allow sufficient categories to capture the diversity of potential effects while maintaining a relatively simple analytical framework. The current set of channels is discussed below:

Employment: all types of formal and 'informal' employment (including self-employment and employment in household enterprises) may be addressed under this heading. Changes will typically impact on cash or kind income flows but other aspects of employment, for example security, status and workloads, may also be considered here. Gender issues will be of considerable importance. Interventions may affect either the aggregate demand for labour or the composition of that demand. Both direct and indirect impacts should be disaggregated by sector, the nature of the explicit or implicit employment contract and associated wage rates in order to assess the potential positive or negative impact on poor households and individuals. In many cases the most important mode of transmission will be indirect. Any intervention which promotes growth in sectors that tend to employ the poor might be expected to generate positive employment impacts for them, at least in the medium term. Growth in the agricultural sector, for example, might be expected to promote a demand both for agricultural labour and for off-farm labour as farming households increase their consumption and investment expenditures. The precise nature of the sector and sub-sector where this growth occurs will clearly influence the employment outcome. For example, a micro-finance project may stimulate growth in small enterprises, leading to increases in both self-employment and wage employment for the unskilled poor, though the latter may be informal, casual and poorly rewarded.

Transfers: this channel is primarily concerned with an examination of the impact of targeted transfers to poor households, either by means of subsidies or direct payments in cash, vouchers or kind. This may be associated with attempts to mitigate the negative impacts of an intervention on the poor, for example the use of exemption schemes to offset increased fees in a reform of the health sector. It can also be used to consider tax payments which might be associated, for example, with the introduction of a compulsory levy or social insurance scheme. Social protection and subsidy schemes should be assessed in terms of their probable specificity – extent to which they reach the intended beneficiaries, and selectivity – extent to which they avoid capture by non-poor groups.

Access: most PRSPs prioritised increased expenditure on health, education, water, sanitation, micro-finance, roads and infrastructure. The associated interventions can be seen in terms of providing or enhancing the access of the poor to goods and services. This may involve the removal of barriers, whether physical or financial, or through improvements to the quality of the goods and services available. Improved access may often usefully be considered in terms of consequent direct and indirect changes in household asset status. Indirect impacts of increased access may be extremely important. One obvious example is that infrastructure investments in roads, grid electricity, telecommunications, etc. will often not only directly improve access to the associated transport, energy and communications services, but thereby indirectly reduce access barriers to services and markets in many other areas. Similarly, availability of affordable financial services may be a crucial determinant of the ability of a poor household to obtain health care for a serious illness. It t is also necessary to distinguish access for the overall population in poor areas and specific sub-sections of that population. Will, for example, women have access to improved health, education or transport services, or do they have additional cultural barriers to overcome? What are the implications of making grid electricity available to the majority of the poor population in a

given area, while the most vulnerable, perhaps because of their inability to meet connection charges or the remote location of their dwellings, are unable to gain access?

Authority: this term is used to address issues relating to formal and informal institutions, organisations, relationships and power structures. The channel would also be used to examine the effects on poor households of changes in political, legal, social or cultural factors. It is seen as particularly important in addressing issues of empowerment, equity and inclusion. The term 'institutions' is used to cover a very wide range of 'rules and regulations' which affect the lives of the poor. It would include, for example, laws governing land rights, civil service codes of conduct and accepted behavioural norms in specific population groups. An 'organisation' can be any group of individuals who come together with a common purpose and operate within an institutional framework – abiding by a set of rules and regulation – which will typically include a structured decision making process and a defined hierarchy of members. Again it will cover a very wide range, from a government ministry to a village school committee. Indirect effects may arise from responsive changes in the behaviour of economic agents which may have considerable consequences for growth and distribution. Given that markets are highly influenced by the power relationships between producers, traders and consumers, the ability or lack of ability of an intervention to change these relationships, for example by increasing the opportunities for genuine competition, may have a substantial impact on the potential poverty reduction impact. Similarly, normative rules governing intra-household production relationships (e.g. do men capture the benefits from increased cash crop production) may play a major role in determining both the production response of households to income generating opportunities and the poverty impact at the individual level.

Assets: there has been considerable work over the last decade on the importance of household assets in determining the available livelihood strategies of poor households. The ability either to cope with adversity or take advantage of opportunities is seen as highly correlated with the extent to which households are in possession of (or have access to) five types of livelihood asset:

- physical (buildings, tools, equipment, livestock, access to infrastructure, etc.)
- natural (land, water, forest, natural resources, etc.)
- human (labour supply, education, skills, knowledge, health, nutritional status, etc.)
- social (networks, groups, relationships) and
- financial (savings, access to credit, pension or similar guaranteed income, etc.).

Interventions which tend to increase or decrease the value of any of these assets will change the livelihood options of poor households in ways which may impact on their welfare. 'Policy reforms can change the context and income-generating potential of assets. Investments can add new assets or increase the efficiency of existing household assets, and also improve households' risk management capacity to protect assets. After all is said and done, a household's asset portfolio will determine whether growth and poverty reduction can be achieved and sustained over time' (Siegal, 2005). Changes in asset holdings will also have consequences in terms of the vulnerability of households to external shocks. For example, increasing the area of irrigated, cultivable land or construction of weather-proof crop storage buildings will tend to improve food security. Road building or maintenance, extension of electricity grids or telecommunications networks, school or hospital construction, may directly result in changes in the value of the land assets of the poor. A range of conservation activities, either by means of technical intervention or regulatory reforms may similar increase or decrease the value of natural resources, for example forests or marine fisheries, possibly with very different impacts for the poor over the short and medium term.

Human assets will clearly be directly impacted through education, training, health and nutrition interventions. These may involve the provision or enhancement of services, or, as discussed above, the removal of barriers which limit access for some poor individuals. They may also be impacted indirectly, for example if enhancement of energy services reduces indoor air pollution and the chronic respiratory disease with which it is associated. Social assets may be directly impacted by interventions or programmes which specifically invite extensive and long-run community involvement, for example the establishment of social funds or micro-finance organisations. The involvement of existing community organisations, for example women's or youth groups, in implementation or monitoring of interventions, possibly with financial, technical or capacity building support may indirectly enhance the status and cohesion of such groups. Finally, the value of financial assets may be directly impacted by interventions which affect the savings and credit environment, for example the establishment of a micro-credit or micro-finance organisation, or programmes which involve the modification of financial sector legislation or regulation. They may be indirectly affected by interventions which either entail the use of financial assets (reduction in savings or accumulation of debt) for participation, for example investment in equipment or inputs, and/or generate sufficient income to allow savings.

It is important to consider what combination of assets might be required for a poor household to benefit from any proposed intervention and to estimate what proportion of households is likely to possess or be able to attain such a combination. For example, expanded production of a cash crop may be an appropriate medium term livelihood strategy – but only if households have the necessary combination of land, knowledge, skills, equipment, and financial assets to allow efficient production, plus the skills and access to communications and transport services necessary for effective marketing. Households which lack the latter are likely to see much of the value added that they have generated benefiting others.

A substantial number of PSIA exercises have been documented and are available on the World Bank website (http://go.worldbank.org/K59U0MA3V0). Here, just one relevant example is considered to illustrate the process. It is one of a number of relatively low-cost, short time-period pilot exercises funded by DFID. Each had a forty day time limitation and involved two international and at least two national consultants. This particular exercise focused on an export promotion initiative by the Uganda government. The component selected for detailed analysis was the provision of improved inputs to farmers on favourable terms and government action to ease access to export markets. The intervention was thus intended to operate via the prices transmission channel. Second round effects were then assumed to be triggered via the employment channel (increased production in response to incentives) and third round effects via the assets channel (increased value of human and/or physical assets as additional income is consumed and/or invested).

The PSIA adopted an approach (Table 8), which questioned each of these assumptions in some detail. In doing so, it can be interpreted as following to some extent the realistic evaluation paradigm described in section 1. The first issue considered was how various types of intended beneficiary would respond to the government supported incentives. The analysts argued that this would depend on the actual and perceived “incentives and constraints facing different categories of direct producer - large and small, obviously, but also men and women, adults and children, smallholders and wage labourers”. Each would have to assess, for example, if the potential gains envisaged from the input subsidies would accrue to them or, perhaps because of the lack of a competitive and efficient marketing chain, to a variety of intermediaries. Similarly they would have to decide to what extent any such gains would be translated in their increased well-being. One key issue identified was the fallacy that all poor farm households would respond as cooperative, income-sharing units. In fact, on the basis of previous evidence and case study exercises the authors concluded that “under typical conditions, women may be expected to withhold their labour or even sabotage cash crops in subtle ways because they know they will not benefit from the income earned.” The validity of their concerns, it was argued, was supported by analysis of large scale surveys which indicated that “Areas that have had sustained success in commercialisation of smallholder production, and otherwise enjoy favourable conditions, seem incapable of achieving equivalent improvements in nutrition and child survival.”

Table 8 : Export promotion: a typology of impact issues

Impact issues	
Incentive issues	Distributional issues
1) The opportunities, constraints and risks to producers and exporters arising from international market conditions: does the policy help to convey these to producers and traders? Example: coffee producers and traders become focused on quality-enhancement.	1) The structure of production: how will this affect the distribution of benefits among firms and households? - scale of operations, - factor intensities, - linkages, - region, urban/rural, etc. Example: predominance of smallholder production helps to spread income effects.
2) Interventions to correct for market failures (e.g. to ensure the environmental sustainability of production): does the policy provide them? Example: rules on net sizes are enforced, preventing over-fishing.	2) Non-market instruments to redistribute benefits (e.g. taxes): are they needed, and if so, are they being created? Example: tax system captures a share of profits from enclave type of export enterprise, and this finances pro-poor public spending.
3) Supply chains: are they sufficiently competitive to transmit incentives to producers? Example: share of world price reaching farm gate is high by international standards, making export crop growing relatively attractive.	3) Supply chains: are they sufficiently competitive to permit an equitable distribution of incomes between levels? Example: all steps in supply chain for exportable fish are free of monopoly and collusion.
4) Intra-household production decision units: is their structure likely to encourage a vigorous supply response? Example: in practice, spouses share land rights, income from cash crops and responsibility for children; so, they respond jointly to price signals.	4) Intra-household distribution of consumption: will increased cash inflows reduce poverty in all dimensions? Example: in practice, spouses share land rights, income from cash crops and responsibility for children; so, cash is used to buy food and protect children's health.

Source: adapted from Booth et al. 2003: 14

5 References

ActionAid (2006) Accountability, Learning and Planning System, www.actionaid.org.uk/doc_lib/alpsfinal2006.pdf (accessed 6 October 2010.

Adam, C.; Chambas, G., Guillaumont, P., Jeanneney, S.G., Gunning, and W.,J. (2004) 'Performance-based conditionality: A European perspective', World Development 32.6: 1059 - 1070

Asian Development Bank (ADB) (2007) Improving Local Governance and Pro-Poor Service Delivery: Citizen Report Card Learning Toolkit, Manila: ADB

Asian Development Bank (ADB) (2002) Economic Analysis of Policy-Based Operations: Key Dimensions, Asian Development Bank, Manila: ADB

Anand, S. and Hanson, K. (1998) 'DALYs: Efficiency versus equity', World Development 26.2: 307 - 310

Aulakh, A.K. and Anand, S.S. (2007) 'Sex and gender subgroup analysis of randomized trials: the need to proceed with caution', Women's Health Issues 17: 342 - 350

Banerjee, A. V. and Duflo, E. (2008) The Experimental Approach to Development Economics
CEPR Discussion Paper (DP7037), Centre for Economic Policy Research, London, UK. www.cepr.org/pubs/new-dps/dplist.asp?dpno=7037 (accessed 10 August 2010)

Barnes, M.; Sullivan, H. and Matka, E. (2003) 'Evidence, Understanding and Complexity: Evaluation in Non-Linear Systems', Evaluation 9: 263 - 282

Bastoe, P.O. (2006) 'Implementing Results-Based Management', in R.C. Rist and N. Stame (eds), From Studies to Streams: Managing Evaluative Systems, London: Transaction Publishers

Bezzi, C. (2006) 'Pragmatic Evaluation', Evaluation 12.1: 56 - 76

Bird, K.; Curran, Z., Evans, A., and Plagerson, S. (2005) What has DFID learned from the PSIA Process?, London: Overseas Development Institute (ODI)

Blamey, A., and Mackenzie, M. (2007) 'Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges?', Evaluation 13: 439 - 455

Blasi, Z.D.; Harkness, E., Ernst, E., Georgiou, A. and Kleijnen, J. (2001) 'Influence of context effects on health outcomes: a systematic review', The Lancet 357.9258: 757 - 762

Booth, D. (2005) Missing Links in the Politics of Development: Learning from the PRSP Experiment, Working Paper 256, London: ODI

Booth, D.; Christiansen, K. and de Renzio, P. (2005) Reconciling Alignment and Performance in Budget-Support Programmes: What Next? Practitioner's Forum on Budget Support, Cape Town: World Bank

Booth, D.; Kasente, D., Mavrotas, G., Mugambe, G. and Muwonge, A. (2003) Poverty and Social Impact Analysis: The strategic exports initiative in Uganda, Oxford Policy Management . Oxford, UK.
http://siteresources.worldbank.org/INTPSIA/Resources/490023-1120841262639/14689_Uganda_Final_PSIA.pdf (accessed 25 June 2010).

Booth, D. and Lucas, H. (2004) 'Monitoring progress towards the Millennium Development Goals at country level', in R. Black and H. White (eds) Targeting development: Critical perspectives on the Millennium Development Goals and International Development Targets, London: Routledge

Bourguignon, F.; Bénassy-Quéré, A., Dercon, S., Estache, A., Gunning, J.W., Kanbur, R., et al. (2008) Millennium Development Goals at Midpoint: Where do we stand and where do we need to go? European Report on Development, Brussels: European Commission

Brousselle, A.; Contandriopoulos, D. and Lemire, M. (2009) 'Using Logic Analysis to Evaluate Knowledge Transfer Initiatives: The Case of the Research Collective on the Organization of Primary Care Services', Evaluation 15.2: 165 - 183

Brugha, R.; Starling, M. and Walt, G. (2002) 'GAVI, the first steps: lessons for the Global Fund', The Lancet 359: 435-438.

Chambers, R. (2008) Revolutions in Development Enquiry, London: Earthscan

Chambers, R. (1997) Whose Reality Counts: Putting the Last First, London: Intermediate Technology Publications

Chambers, R. (1986) 'Rapid rural appraisal: rationale and repertoire', Public Administration and Development 1.2: 95

Connell, J.; Kubisch, A., Schorr, L. and Weiss, C. (eds) (1995) New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts, Washington, DC: Aspen Institute

Coudouel, A.; Dani, A.A. and Paternostro, S. (eds) (2006) Poverty and Social Impact Analysis of Reforms: Lessons and Examples from Implementation, Washington, DC: World Bank

Council for International Organizations of Medical Sciences (CIOMS) (2002) International Ethical Guidelines for Biomedical Research Involving Human Subjects, Geneva: CIOMS

Davies, P. and Boruch, R. (2001) 'The Campbell Collection', BMJ 323: 294 - 295

Davies, R. (2005) 'Scale, Complexity and the Representation of Theories of Change', Evaluation 11.2: 133 - 149

Deaton, A. (2009) 'Randomization in the tropics, and the search for the elusive keys to economic development', The Keynes Lecture, London: British Academy

Dehejia, R.H. and Wahba, S. (2002) 'Propensity Score-Matching Methods for Nonexperimental Causal Studies', Review of Economics and Statistics 84.1: 151 - 161

Department for International Development (DFID) (1999) 'The Livelihoods Framework', Sustainable Livelihoods Guidance Sheets 2, London: DFID

Dixon-Woods, M.; Agarwal, Young, B., Jones, D. and Sutton, A. (2004) Integrative approaches to qualitative and quantitative evidence, London: NHS Health Development Agency

Duflo, E.; Dupas, P. and Kremer, M. (2008) Peer effects and the impact of tracking: evidence from a randomized evaluation in Kenya, National Bureau of Economic Research (NBER)Working Paper Series, Cambridge, MA: NBER

Earle, K. (ed) (2004) Creativity and Constraint: Grassroots Monitoring and Evaluation and the International Aid Arena, NGO Management and Policy Series: INTRAC, Oxford, UK.

Eldridge, S.; Ashby, D., Bennett, C., Wakelin, M. and Feder, G. (2008) 'Internal and external validity of cluster randomised trials: systematic review of recent trials', British Medical Journal 336.7649: 876 - 880

Ernst, E.; Pittler, M.H., Wider, B. and Boddy, K. (2008) Oxford Handbook of Complementary Medicine, Oxford: Oxford University Press

EuropeAid Cooperation Office (2004) Project Cycle Management Guidelines, Brussels: European Commission (EC)

European Commission (2007) Evalsed: the resource for the evaluation of socio-economic development. http://ec.europa.eu/regional_policy/sources/docgener/evaluation/evalsed/index_en.htm (accessed 10 June, 2010.

Feachem, R. G. A. and Sabot, O. J. (2006) 'An examination of the Global Fund at 5 years', The Lancet 368

Fulbright-Anderson, K.; Kubisch, A. and Connell, J. (eds) (1998) New Approaches to Evaluating Community Initiatives Volume 2: Theory, Measurement, and Analysis, Washington, DC: Aspen Institute

Geddes, M. (2006) National evaluation of local strategic partnerships: theories of change issues paper, London: Department for Communities and Local Government

Gilson, L.; Kalyalyab, D., Kuchlerc, F., Lakea, S., Orangad, H. and Ouendoe, M. (2001) 'Strategies for promoting equity: experience with community financing in three African countries', Health Policy 58.1: 37 - 67

Giordano, R.; Passarella, G., Uricchio, V. F. and Vurro, M. (2007) 'Integrating conflict analysis and consensus reaching in a decision support system for water resource management', Journal of Environmental Management 84: 213 - 228

Global Fund (2008) Data Quality Audit Tool, : Global Fund. www.theglobalfund.org/documents/monitoring_evaluation/ME_RDQA_Guidelines_en/ (accessed 6 July 2010)

Global Fund (2008) Monitoring and Evaluation Manual 1: Performance-Based Funding and M&E in Practice, : Global Fund. www.theglobalfund.org/documents/monitoring_evaluation/ME_Module03_Manual_en/ (accessed 6 July 2010).

Global Fund (2007) M&E Plan Guidelines: Global Fund. www.theglobalfund.org/en/me/documents/ (accessed 6 July 2010).

Global Fund (2006) Monitoring and Evaluation Toolkit: HIV/AIDS, Tuberculosis and Malaria: Global Fund. www.theglobalfund.org/documents/monitoring_evaluation/ME_MonitoringEvaluation_Toolkit_en/ (accessed 6 July 2010.

Grosskurth, H.; Gray, R., Hayes, R., Mabey, D. and Wawer, M. (2000) 'Control of sexually transmitted diseases for HIV-1 prevention: understanding the implications of the Mwanza and Rakai trials', Lancet 355: 1981 - 1987

GTZ (undated) 'Practitioner's guide: Consultative Impact Monitoring of Policy (CoIMPact)', Method Finder: GTZ. www.methodfinder.net/pdfmethods/gtz/example/gtz_example41_1.pdf (accessed 6 August 2010).

Guba, E.G. and Lincoln, Y.S. (1989) Forth Generation Evaluation, Newbury Park, CA: Sage

Habicht, J.P.; Victora, C.G. and Vaughan, J.P. (1999) 'Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact', Int. J. Epidemiol 28.1: 10 – 18

Haddad, L.; Lindstrom, J. and Pinto, Y. (2010) 'The Sorry State of M&E in Agriculture: Can People-Centred Approaches Help?', IDS Bulletin 41.6, IDS: Brighton

Hansen, H.F. and Rieper, O. (2009) 'The Evidence Movement: The Development and Consequences of Methodologies in Review Practices', Evaluation 15.2: 141 - 163

Hayes, L. (2005) Open on impact: Slow progress in World Bank and IMF poverty analysis, EURODAD. www.eurodad.org/uploadedFiles/Whats_New/Reports/PSIA_webFINAL.pdf (accessed 6 June 2010).

Higgins, J.P. and Green, S. (2008) Cochran Handbook for Systematic Reviews of Interventions Version 5.0.0. Cochran Collaboration. http://www.cochrane.org/training/cochrane-handbook. (accessed 10 June 2010.

Holland, J. (ed) (2007) Tools for Institutional, Political, and Social Analysis of Policy Reform A Sourcebook for Development Practitioners, Washington, DC: World Bank

Holvoet, N. and Renard, R. (2007) 'Monitoring and evaluation under the PRSP: Solid rock or quicksand?', Evaluation and Program Planning 30: 66 - 81

Health Partners International (HPI) (2009) Peer and Participatory Rapid Health Appraisal for Action (PPRHAA), Lewes, UK: HPI

INGO (2005) Accountability Charter, Berlin: Berlin Civil Society Center. www.ingoaccountabilitycharter.org/wpcms/wp-content/uploads/INGO-Accountability-Charter_logo1.pdf (accessed 12 June 2010).

International Bank for Reconstruction and Development (IBRD) (2007) Tools for Institutional, Political and Social Analysis of Policy Reform: A Sourcebook for Development Practitioners, Washington, DC: World Bank

Ioannidis, J.P. (2005) 'Contradicted and initially stronger effects in highly cited clinical research', JAMA 294.2: 218 - 228

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988) 'Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction', Lancet(ii): 349 - 360

Kanbur, R. (ed) (2003) Q-Squared : Combining Qualitative and Quantitative Methods in Poverty Appraisal, Delhi: Permanent Black

Kapiriri, L.; Robbestad, B. and Norheim, O.F. (2003) 'The relationship between prevention of mother to child transmission of HIV and stakeholder decision making in Uganda: implications for health policy', Health Policy 6: 199 - 221

Kaplan, R.S. and Norton, D.P. (1992) 'The balanced scorecard: measures that drive performance', Harvard Business Review Jan-Feb: 71 - 80

Klasen, S. (2005) Economic Growth and Poverty Reduction: Measurement and Policy Issues, Working Paper 246, Paris: OECD Development Centre

Koenig, G. (2009) 'Realistic Evaluation and Case Studies: Stretching the Potential', Evaluation 15: 9 - 30

Kremer, M. and Holla, A. (2008) Pricing and Access, Lessons from randomized evaluations in education and health, Department of Economics, Cambridge, MA: Harvard University

Lancet (2004) The World Bank is finally embracing science, The Lancet, 364: 731-732.

LATH Consortium (2001) Immunization data quality audit evaluation report: final report from Deloitte Touche Tohmatsu Emerging Markets, Euro Health Group, and Liverpool Associates in Tropical Health, Geneva: World Health Organization (WHO)

Lay, M. and Papadopoulos, I. (2007) An Exploration of Fourth Generation Evaluation in Practice Evaluation 13.4: 495 - 504

Lindstrom, J. (2009): What is the state of M&E in agriculture? Findings of the ALINe online consultation survey, October 2009, IDS

Lockheed, M.E. (2009) 'Evaluating Development Learning: The World Bank Experience', Evaluation 15: 113 - 126

Lucas, H.; Yang, H., Zhang, T. and Lin, V. (2008) 'Monitoring and Evaluation as Tools for Policy', in V. Lin, Y. Guo, D. Legge and Q. Wu (eds) Health Policy in Transition: The Challenges for China, Beijing: Peking University Medical Press

Mackenzie, M. and Blamey, A. (2005) 'The Practice and the Theory: Lessons from the Application of a Theories of Change Approach', Evaluation 11.2: 151 - 168

McNamara, P. (2006) 'Provider-specific report cards: a tool for health sector accountability in developing countries', Health Policy and Planning 21: 101 - 109

Meja, V. (2006) Linking PRSPs to MDGs: Some Macroeconomic Policy Options for Sub Saharan Africa, Occasional Papers: AFRODAD

Millstone, E., Van Zwanenberg, P. and Marshall, F. (2010) 'Monitoring and Evaluating Agricultural Science and Technology Projects: Theories, Practices and Problems', IDS Bulletin 41.6, IDS: Brighton

Milne, L.; Scotland, G., Tagiyeva-Milne, N. and Hussein, J. (2004) 'Safe Motherhood Program Evaluation: Theory and Practice', Journal of Midwifery & Women's Health 49.4: 338-344 .

Mortimer, D.; Segal, L (2008). Comparing the Incomparable? A Systematic Review of Competing Techniques for Converting Descriptive Measures of Health Status into QALY-Weights. Medical Decision Making 28 (1): 66–89.

Mosse, D. (1998) 'Process-oriented approaches to development practice and social research', in D. Mosse, J. Farrington and A. Rew (eds) Development as Process: Concepts and Methods for Working with Complexity, London: Routeledge/ODI

Murray, C.J.L.; Lopez, A.D. and Wibulpolprasert, S. (2004) 'Monitoring global health: time for new solutions', BMJ 329.7474: 1096 - 1100

National Institute for Health and Clinical Excellence (NICE) (2009) Methods for the development of NICE public health guidance (second edition), London: NICE

Nielsen, S.B. and Ejler, N. (2008) 'Improving Performance?: Exploring the Complementarities between Evaluation and Performance Management', Evaluation 14: 171 - 192

Norton, A. (2001) A rough guide to PPAs: Participatory Poverty Assessment an introduction to theory and practice, London: ODI

OECD DAC (2007) Promoting pro-poor growth: a practical guide to Ex Ante Poverty Impact Assessment, DAC Guidelines and Reference Series, Paris: OECD

OECD DAC (2005) Paris Declaration on Aid Effectiveness: Ownership, Harmonisation, Alignment, Results and Mutual Accountability, Paris: OECD

OXFAM (undated) Oxfam GB Evaluation Guidelines, Oxford, UK: OXFAM. www.oxfam.org.uk/resources/evaluations/downloads/oxfam_evaluation_guidelines.pdf. (accessed 12 June 2010).

OXFAM (2002) Influencing Poverty Reduction Strategies: A guide, Oxford, UK: OXFAM. www.oxfam.org.uk/what_we_do/issues/democracy_rights/downloads/prsp_guide.pdf (accessed 12 June 2010).

Pain, C. and Kirsch, R. (2002) Document review on the challenges in monitoring the PRSP, Beyond the review: sustainable development and PRSP, Berlin: GTZ

Parkinson, S. (2009) 'Power and perceptions in participatory monitoring and evaluation', Evaluation and Program Planning 32: 229 - 237

Patton, M. Q. (2002) Utilsation-focused evaluation - a checklist, The Evaluation Center, Kalamazoo, MI: Western Michigan University

Patton, M. Q. (1997) Utilisation-focused evaluation: The new century text, Thousand Oaks, CA: Sage

Patton, M. Q. (1990) 'The evaluator's responsibility for utilisation', in M. Alkin (ed) Debates on Evaluation, Newbury Park, CA: Sage

Pawson, R. (2006) Evidence based policy: a realist perspective, London: Sage

Pawson, R (2003) 'Nothing as Practical as a Good Theory', Evaluation 9.4: 471 - 490

Pawson, R. and Tilley, N. (1998) 'Caring Communities, Paradigm Polemics, Design Debates', Evaluation 4.1: 73 - 90

Pawson, R. and Tilley, N. (1997) Realistic Evaluation, London: Sage

Pawson, R.; Greehaigh, T., Harvey, G. and Walshe, K. (2005) 'Realist review - a new method of systematic review for complex policy interventions', Journal of Health Services Research and Policy 10 Supplement 1: 21 - 33

Pereira da Silva, L.A. and Bourguignon, F. (eds) (2003) Toolkit for Evaluating the Poverty and Distributional Impact of Economic Policies, Washington, DC: World Bank

Pereira da Silva, L.A.; Bourguignon, F. and Bussolo, M. (2008) The Impact of Macro-Economic Policies on Poverty and Income Distribution: Macro-Micro Evaluation Techniques and Tools, Washington, DC: World Bank

Peters, D.H.; Noor, A.A., Singh, L.P., Kakar, F.K., Hansena, P.M. and Burnhama, G. (2007) 'A balanced scorecard for health services in Afghanistan', Bulletin of the World Health Organization 85: 146 - 151

Peters, D.H.; Sameh El-Saharty, S., Siadat, B., Janovsky, K. and Vujicic, M. (2009) Improving Health Service Delivery in Developing Countries: From Evidence to Action, Washington, DC: World Bank

Prennushi, G.; Rubio, G. and Subbarao, K. (2002) 'Monitoring and Evaluation', in J. Klugman (ed) A Sourcebook for Poverty Reduction Strategies, Washington, DC: World Bank

Radelet, S. and Siddiqi, B. (2007) 'Global Fund grant programmes: an analysis of evaluation scores', The Lancet 369: 1807-1813.

Ravallion, M. (2002) 'The mystery of the vanishing benefits: an introduction to impact evaluation', World Bank Economic Review 15(1):115-140.

Renard, R. (2006) The cracks in the new aid paradigm, Discussion paper 2006/1, Antwerp: Institute of Development Policy and Management, University of Antwerp

Rist, R.C. (2006) 'The "E" in Monitoring and Evaluation: Using Evaluative Knowledge to Support a Results-Based Management System', in R.C. Rist and N. Stame (eds) From Studies to Streams: Managing Evaluative Systems, London: Transaction Publishers

Robb, C. (2002) Can the poor influence policy: Participatory policy assessments in the developing world, Washington, DC: World Bank

Rogers, P.J. (2008) 'Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions', Evaluation 14: 29 - 28

Sabates-Wheeler, R.; Butters, S. and Greeley, M. (2010) Risk and Agricultural Livelihoods: How does project design incorporate and influence farm-level risk? Research Report I. Agriculture Learning and Impact Network (ALINe) March 2010.

Schnell, S. and Forster, R. (2003) Participation in Monitoring and Evaluation of PRSPs A Document Review of Trends and Approaches Emerging from 21 Full PRSPs, Social Development Papers, Washington, DC: Social Development Department, World Bank

Scriven, M. (2010) 'A summative evaluation of RCT methodology: and an alternative approach to causal research', Journal of Multidisciplinary Evaluation 5.9: 11 - 24

Secker, J.; Bowers, H., Webb, D. and Llanes, M. (2005) 'Theories of change: what works in improving health in mid-life?', Health Education Research Theory & Practice 20.4: 392 - 401

Shadish, W.R.; Cook, T.D. and Campbell, D.T. (2001) Experimental and Quasi-experimental Designs for Generalised Causal Inference, Boston, New York: Houghton Mifflin

Sherry, J.; Mookherji, S. and Ryan, L. (2009) Five-Year Evaluation of the Global Fund to Fight AIDS, TB and Malaria, Synthesis of study areas 1, 2 and 3, Geneva: Global Fund to fight AIDS, Malaria and Tuberculosis

Siegel, P. (2005) Using an Asset-Based Approach to Identify Drivers of Sustainable Rural Growth and Poverty Reduction in Central America: A Conceptual Framework, World Bank Policy Research Working Paper 3475, Washington, DC: World Bank

Ssewakiryanga, R. (2005) 'Experience of Uganda's PPA in implementing and monitoring poverty reduction', in A. Hughes and N. Atampugre (eds) Civil society and poverty reduction, London: International Institute for Environment and Development (IIED)

Stame, N. (2004) Theory-Based Evaluation and Varieties of Complexity, Evaluation 10.1: 58 - 76

Stern, E. (2009) 'Editorial', Evaluation 15.5: 5 - 7

Stern, E. (2008a) 'Editorial', Evaluation 14: 267 - 269

Stern, E. (2008b) 'Evaluation: Critical for Whom and Connected to What?', Evaluation 14: 249 - 257

Stern, E. (2004) 'What Shapes European Evaluation? A Personal Reflection', Evaluation 10.1: 7 - 15

Thin, N.; Underwood, M. and Gilling, J. (2001) Sub-Saharan Africa's Poverty Reduction Strategy Papers from Social Policy and Sustainable Livelihoods Perspectives, London: Oxford Policy Management/DFID

Tilley, N. (2000) Realistic Evaluation: An Overview, Founding Conference of the Danish Evaluation Society, September 2000.
www.evidence-basedmanagement.com/research_practice/articles/nick_tilley.pdf
(accessed 15 May 2010).

United Nations Development Program (UNDP)/World Bank (2002) How do the Millennium Development Goals Relate to Poverty Reduction Strategy Papers? , New York: UNDP

Van der Knaap, P. (2004) 'Thory-based evaluation and learning: possibilities and challenges', Evaluation 10.1: 16 - 34

Van der Knaap, P. (1995) 'Policy Evaluation and Learning: Feedback, Enlightenment or Argumentation?', Evaluation 1.2: 189 - 216

Victora, C.G.; Habicht, J.P. and Bryce, J. (2004) 'Evidence-Based Public Health: Moving Beyond Randomized Trials', Am J Public Health 94.3: 400 - 405

Virtanen, P. and Uusikyla, P. (2004) 'Exploring the Missing Links between Cause and Effect: A Conceptual Framework for Understanding Micro–Macro Conversions in Programme
Evaluation', Evaluation 10.1: 77 - 91

Wallace, T. (1997) 'New Development Agendas: Changes in UK NGO Policies and Procedures', Review of African Political Economy 24.71: 35 - 55

Weiss, C. (2000) 'Which links in which theories shall we evaluate?', in P. Rogers, T. Hacsi, A. Petrosino and T. Huebner (eds) Program Theory in Evaluation: Challenges and Opportunities, San Francisco, CA: Jossey-Bass

Weiss, C. (1998) Evaluation: Methods for Studying Programs and Policies, Upper Saddle River, NJ: Prentice Hall

Weiss, C. (1997) 'Theory-Based Evaluation: Past, Present and Future', New Directions for Program Evaluation 76: 41 - 55

Weiss, C. (1995) 'Nothing As Practical As Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families', in J. Connell, A. Kubisch, C.L.
Schorr and C. Weiss (eds) New Approaches to evaluating community initiatives: Concepts, Methods, and Contexts, Washington, DC: The Aspen Institute

Weiss, C. (1990) 'Evaluations for Decisions', in Alkin, M (eds) Debates on Evaluation, Newbury Park, CA: Sage

Weiss, C. (1972) Evaluation Research: Methods Of Assessing Program Effectiveness, Englewood Cliffs, NJ: Prentice Hall

Weitzman, B.C.; Silver, D. and Dillman, K.N. (2002) 'Integrating a Comparison Group Design into a Theory of Change Evaluation: The Case of the Urban Health Initiative', American Journal of Evaluation 23.4: 371 - 385

White, H. (undated) Impact evaluation: the experience of the Independent Evaluation Group of the World Bank, Washington, DC: World Bank

White, H. (2007) Evaluating Aid Impact, UNU-WIDER Research Papers 75, Helsinki, Finland, United Nations University – World Institute for Development Economics Research.

www.wider.unu.edu/publications/working-papers/research-papers/2007/en_GB/rp2007-75/ (accessed 12 June 2010).

White, H. (2005a) Challenges in evaluating development effectiveness, IDS Working Paper 242, Brighton: IDS

White, H. (2005b) The road to nowhere? Results based management in international cooperation, in Why did the Chicken cross the road and other stories on development evaluation. KIT Bulletin series 363. Royal Tropical Institute (KIT), KIT Publishers Amsterdam, The Netherlands 71-76. www.kitpublishers.nl/net/KIT_Publicaties_output/ShowFile2.aspx?e=1005 (accessed 12 June 2010).

Wilson, V. and McCormack, B. (2006) 'Critical realism as emancipatory action: the case for realistic evaluation in practice development', Nursing Philosophy 7.1: 45 - 57

Wittes, J. (2009) 'On looking at subgroups', Circulation 912 - 915

Wood, A. (2005) Beyond data: A panorama of CSO experiences with PRSP and HIPC monitoring, Den Haag: Cordaid

World Bank (2008) Good practice note: using poverty and social impact analysis to support development policy operations, Washington, DC: World Bank

World Bank (2005) 'The Community Score Card Process in Gambia', Social Development Note No. 100, Washington, DC: World Bank

World Bank (2003) A User's Guide to Poverty and Social Impact Analysis, Washington, DC: World Bank

World Bank (1999) CWIQ: Core Welfare Indicators Survey Handbook, Washington, DC: World Bank

World Bank/IMF (1999) Poverty Reduction Strategy Papers—Operational Issues, Washington, DC: World Bank/IMF

Zwarenstein, M. and Treweek, S. (2009) 'What kind of randomized trials do we need?', Journal of Clinical Epidemiology 62: 461 - 463

ALINe

PEOPLE-CENTRED PERFORMANCE

ALINe (Agricultural Learning and Impacts Network)
3 Dean Trench Street
London SW1P 3HB
T: +44 207 148 0914 **E:** info@aline.org.uk **W:** www.aline.org.uk