Evaluation: Why, for Whom and How?

Henry Lucas and Richard Longhurst

Abstract The overall objective of this article is to discuss current theoretical debates in the evaluation literature to assess their relevance for agriculture. After noting the 'bad press' of monitoring and evaluation (M&E) in agriculture, the literature is selectively reviewed to see what light is shed from different evaluation paradigms and methodologies. Experiences in the health sector are then examined in greater depth, emphasising debates around randomised controlled trials (RCTs). The final section considers some possible ways forward.

1 The 'bad press' for agriculture M&E: is it justified?

The characterisation of the existing state of M&E in agricultural interventions by those participating in the ALINe consultation survey (Lindstrom 2009) 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 Millstone *et al.*, this *IDS Bulletin*). These include:

- a lack of clarity as to primary objectives – projects typically have multiple objectives entailing complex trade-offs;
- long 'causal chains', in terms of both number of links and overall project duration; 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? In spite of having an apparently coherent objective, 'improving population health status', 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

IDS Bulletin Volume 41 Number 6 November 2010
Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA

Table 1 **Characteristics of different approaches to evaluation**

Perspective	Epistemology	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	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).

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 implementation theory for these interventions is reasonably well defined, and plausible, relatively short, causal chains have been specified. However, this has provided no guarantee of success. For example, progress in reducing maternal and neonatal 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. As discussed in the next section, this has led many to question some of the assumptions underlying mainstream evaluation work in the health sector, which has broadly adopted the 'experimental' paradigm, exemplified by the randomised controlled trial (RCT). In particular, there has been a recent focus on 'programme theory' – attempting to understand how specific types of individual respond to different aspects of an intervention within a specific context.

2 Competing approaches to evaluation

The debate as to which methodologies can best describe and attribute causality in evaluating interventions has been called 'the causal wars' (Scriven 2010; Stern 2008). Milne *et al.* (2004) categorise the contenders under five headings (Table 1).

The 'pragmatic' approach regards evaluators as contracted technicians meeting the needs of their client. Their role is to help that client 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' paradigm (Lay and Papadopoulos 2007), which aims at a 'negotiated settlement' between all stakeholders, attempting to reconcile their diverse perceptions. 'Experimental' evaluators see their task as identifying cause and effect relationships using controlled trials, while those adopting the 'theories of change' approach insist on theoretical explanations of those relationships. Finally, the 'pluralists' seek ways to draw on all these different perspectives and are usually condemned as unprincipled eclectics.

More usefully for present purposes, the European Commission (EC) guide to the evaluation of socioeconomic development initiatives takes a strictly practical approach to the choice of methodology, describing five distinct purposes which may be given priority in the commissioning or implementation of a given evaluation (European Commission 2007):

1 Planning/efficiency: ensuring justification for a policy/programme 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 the performance of programmes and the effectiveness of how they are delivered and managed.
4 Knowledge production: increasing understanding of what works in what circumstances 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.

These different concerns will typically influence methodological preferences.

- *Planning and efficiency* issues may be seen as best approached thorough various forms of impact and cost-benefit analysis, possibly linked to the Logical Framework Approach (EuropeAid 2004).
- Those most concerned with *accountability* will tend to focus on the assessment of performance against agreed targets, with emphasis on quantitative techniques, typically 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 that can be used for intervention management, possibly combined with analysis of factors determining the performance of administrative and service delivery units.
- 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 projects (Peters *et al.* 2009) emphasises the 'knowledge production' objective where the aim is to replicate successful interventions. It argues that: 'the most important lesson from this enquiry is not about *what should be done* to improve health services, but learning about *how to use knowledge* to improve health services'. Demonstrating that an intervention has improved health outcomes is typically much easier but much less useful than explaining how to do it again. In terms of this knowledge production agenda, it may be useful to characterise three widely held theoretical positions:

1 'Experimental' (RCT): Randomised controlled trials provide the only scientific approach to the evaluation of an intervention. If it is not possible to undertake such trials, the RCT benchmark should be approximated as closely as possible by very careful construction of a counterfactual. However, this is very much a second-best option.
2 'Theories of Change' (ToC): It is essential to focus not only on *whether* an intervention succeeded or failed but *why*. By devoting sufficient resources to developing a shared understanding as to how an intervention is intended to work, M&E systems can be designed that will allow us to evaluate the extent to which 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' (RE): The interventions under review are typically complex and dynamic, with a diversity of components adapting to local contexts. Those who participate in the implementation process, including intended beneficiaries and intervention managers, will have a wide range

Table 2 **Judgements on health interventions and implications for evaluation**

Type of judgement	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 to be effective	Is any measured effect on health services or health status likely to be due to the strategy rather than other influences?	Plausibility	Concurrent, non-randomised clusters where intervention is implemented compared to where it is 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).

of characteristics, perceptions and attitudes that shape their responses to these components. Placebo effects – positive or negative responses to the *fact* of the intervention – will typically be large and uncontrollable. The external environment within which the intervention is made will also inevitably give rise to unforeseen effects that vary over the intervention period. Given this reality, it is essentially *irrational* to seek for evidence that given types of intervention 'work'. The use of RCT or 'quasi-experimental' designs is a waste of time and resources in terms of systematic learning. The aim should be to identify the most interesting facets of the intervention ('mechanisms') and explore how they have performed in relation to specific groups of individuals. This will allow the construction of programme theories that genuinely advance our knowledge and can be used to modify the current intervention or design of the next.

One underlying distinction between these three positions relates to the different weights explicitly or implicitly attached to the various evaluation objectives listed above. Thus the proponents of RE tend to focus almost exclusively on the need for systematic learning, rarely addressing issues of accountability. The extent to which a specific intervention has 'succeeded' or 'failed' (or made efficient use of resources) is of limited interest, given that it cannot be seen as providing reliable insights as to the outcome of future similar interventions. By contrast, those advocating the experimental approach are typically very much concerned with these issues.[1]

There is often an underlying assumption that 'experimental' evaluation designs are more 'scientific'.[2] An editorial in the *Lancet* (2004) applauding the increased attention given to RCTs for programme evaluation by the World Bank was entitled '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',[3] a not-for-profit organisation based in the USA 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'. 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'.

A contrary position is adopted by theWorld Bank publication cited above (Peters *et al.* 2009), which reviews a large number of health sector interventions. This argues that different types of 'scientific' evidence are required depending on the objectives of an evaluation, and identifies at least four types of judgement about an intervention that policymakers may wish to make (Table 2).

Where the aim is simply 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 often be appropriate. Attribution is typically assumed in these cases and a simple before-and-after (or preferably time-series) study will suffice. Where there is concern that external factors may have confounded the apparent relationship between intervention and outcomes, a 'plausibility' argument (Habicht *et al.* 1999; Victora *et al.* 2004) may be required. This implies a need for the use of some form of comparator groups to construct a counterfactual – what would have happened without the intervention? Finally, if those commissioning the evaluation demand statistically valid, confidence limited estimates of differences between indicators of change for intervention and non-intervention sites, probability inference based on RCTs is required, though the review also 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).

Few would dispute the advantages provided by genuine RCTs in terms of determining the impact of a given intervention. They provide the same type of 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 have proved one of the most powerful scientific tools available, repeatedly contradicting long-held beliefs based on observational and epidemiological studies. In one meta-analysis, Ioannidis (2005) found that five out of every six findings of such studies could not be replicated. One eminent critic of RCTs argues 'in ideal circumstances, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project.' His concern is that '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' (Deaton 2009).

This is a very common criticism of RCTs, but practically, 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 (or seemed to work), 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 absence of theory becomes important when, as is often the case, findings from repeated studies are *inconsistent*. There is a conclusion here for agriculture: a long run of RCTs in varied contexts showing positive gains for all farmers given access to a new technology, as compared to those who had not, would seem a perfectly rational basis to encourage the spread of that technology, even if we were not entirely sure how those gains had been achieved.

There are two serious concerns with the experimental approach. First, the desire to claim 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 will involve cluster RCTs, which randomise at the level of geographical areas such

as villages or districts, with resource constraints often leading to a small number of large clusters, the worst possible combination from a statistical perspective and one which almost invariably results in very large theoretical sampling errors that may well (if properly estimated!) undermine the credibility of any findings. More problematic are the wide variety of 'quasi-experimental' designs, whose authors sometimes seem to assume that the simple adoption of a method such as propensity score matching (Ravallion 2002) absolves them from critical analysis and interpretation of their 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. There seems no reason to believe that the impact of an agricultural project will not be influenced (positively or negatively) by the responses of participants to their assignment as members of the treatment or control group, irrespective of the substance of that project. Similarly, enthusiastic project managers may often advance the situation of the experimental group, even where this involves going beyond the project terms of reference.

The second area of concern can perhaps be expressed simply as 'the implementation *is* the intervention'. As noted above, the implementation of even an apparently simple *technical* intervention in health or in agriculture typically involves a complex *social* project. However detailed the project theory of change, the interaction of its components with the diverse perceptions and attitudes of the target population and other stakeholders will almost invariably 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, the establishment of 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 may 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. This should raise doubts as to the likely outcome if the exercise were repeated with a different target population.

The proponents of ToC and RE 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. For a ToC evaluator the response should 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.

3 Combined methods?

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.

An alternative approach, probably acceptable to RE advocates, would be to address accountability using what some have advocated as an alternative to traditional evaluation, results-based monitoring and evaluation (RBME) (Nielsen and Ejler 2008). This 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' (Nielsen and Ejler 2008: 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: 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.

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 (performance-based monitoring), is an interesting possibility. The absence of a control group may be seen as a serious objection by some. However, many policymakers 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.

In a discussion of the 'paradigm wars' being waged between proponents of different approaches to evaluation, Pawson and Tilley (1998) propose that the most productive outcome 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 priorities of different stakeholders – 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. As an evaluation will involve trade-offs between objectives, the need in terms of evaluation design is for transparency in setting out objectives and a realistic assessment of the extent to which each can be achieved.

Notes

1 The first principle of the influential International Initiative for Impact Evaluation (3ie) 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' www.3ieimpact.org (accessed 16 August 2010).
2 Note that the use of RCTs in the physical and biological sciences is in practice limited to a very narrow range of activities, mainly testing the efficacy and safety of drugs and foodstuffs.
3 www.coalition4evidence.org (accessed 16 August 2010).

References

Deaton, A. (2009) 'Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development', The Keynes Lecture, London: British Academy

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

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 16 August 2010)

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

Ioannidis, J.P. (2005) 'Contradicted and Initially Stronger Effects in Highly Cited Clinical Research', *JAMA* 294.2: 218–28

The Lancet (2004) 'The World Bank is Finally Embracing Science', *The Lancet* 364

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, IDS

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

Mortimer, Duncan and 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

Nielsen, S.B. and Ejler, N. (2008) 'Improving Performance?: Exploring the Complementarities between Evaluation and Performance Management', *Evaluation* 14.171–92

Patton, M.Q. (2002) *Utilization-focused Evaluation – A Checklist*, The Evaluation Center, Western Michigan University

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

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

Ravallion, M. (2002) 'The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation', *World Bank Economic Review* 15.115

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: 2–22

Sabates-Wheeler, Rachel; Butters, Saul and Greeley, Martin (2010) *Risk and Agricultural Livelihoods: How does Project Design Incorporate and Influence Farm-level Risk?*, Research Report I, Agriculture Learning and Impact Network (ALINe)

Scriven, M. (2010) 'A Summative Evaluation of RCT Methodology: And an Alternative Approach to Causal Research', *Journal of Multidisciplinary Evaluation* 5.9: 11–24

Stern, E. (2008) 'Evaluation: Critical for Whom and Connected to What?', *Evaluation* 14: 249–57

Victora, C.G.; Habicht, J.-P. and Bryce, J. (2004) 'Evidence-based Public Health: Moving Beyond Randomized Trials', *Am J Public Health* 94.3: 400–5

Weiss, C. (1995) 'Nothing as Practical as Good Theory: Exploring Theory-based Evaluation for Comprehensive Community Initiatives for Children and Families', in J.P. Connell, A.C. Kubisch, L.B. Schorr and C.H. Weiss (eds), *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*, Washington DC: The Aspen Institute: 65–92