KIT

KIT WORKING PAPERS 2015-3

Impact assessment and the quest for the Holy Grail

Roger Bymolt

r.bymolt@kit.nl

Evaluation is seen as vital for both accountability and learning purposes. This involves understanding not only *what* worked but also the process of change and *why* and *how* an intervention worked. Donors, programme managers and evaluators often claim to seek not only successful outcomes, but the 'holy grail' of impact. This paper surveys the minefield of what impact is and how it can be reliably assessed, from the perspectives of proponents favouring (quasi)experimental, quantitative designs to those oriented towards the qualitative. It is noted that most programmes do not have sufficient budgetery resources for large scale (quasi)experimental designs and on the other hand purely qualitative designs often lack the rigour or sample sizes to accurately measure impact, even if they are often good at describing processes and perceptions of change. The paper discusses why mixed-methods approaches can be a good option for evaluating many programmes, depending on what is being mixed. The paper concludes that 'impact assessment' is not a term confined to a particular design or method, but that it is most important to apply a design which is appropriate to the intervention, and of course – that elephant in the room – the budget.

Keywords: Impact assessment; quantitative survey; qualitative data collection; PADev; mixed methods; experimental design

Photo: Roger Bymolt

Conducting focus group discussions in Lake Naivasha

Introduction

Institutionalising evaluation in the development sector is an impressive achievement. Large multilateral organisations right down to local NGOs have all come to recognise the importance of evaluation, both as an accountability tool, and for learning purposes. Whilst a certain degree of learning comes from finding out *what* worked, there is a deeper desire to understand the process of change behind *why* and *how* an intervention worked. Lessons learned can be applied when scaling up an intervention in a given area, or to enhance the chances of success when broadening to other contexts. Evaluations contribute to the body of evidence that donors, policy makers, practitioners and academics all draw on to (hopefully) inform their decision making (IPEN/UNICEF, 2006; DFID 2012).

A number of factors have led to evaluation being given increased attention over the past decade or so. These include the desire to track and report on progress towards achieving the Millennium Development Goals, The Paris Declaration on Aid Effectiveness, the Evidence Based Policy movement, and generally a stronger awareness among policy makers and programmers as to its utility. Evaluation is far from being a project completion exercise for its own sake. In fact, EvalPartners, the global movement to strengthen national evaluation capacities, has declared 2015 as International Year of Evaluation (see http://mymande.org/evalyear).

Nevertheless, while we all recognise its importance, policy makers, programming staff and academics alike have had trouble when it comes to judging the reliability of a study's findings, and the extent to which lessons learned can be generalised to other contexts and put into use there (CDI 2013). This piece examines the prevailing thinking and trends in evaluation (specifically impact assessment), which is important for understanding how a contest of ideas has become something of a battleground in the quest for better evidence. It is particularly relevant for the many organisations, including CFC, who desire methodological rigour in their evaluations and yet have to be realistic when it comes to evaluation budgets.

Photo: Roger Bymolt

Conducting surveys in Ethiopia using digital tablets, for the CFC project 'Wealth creation through integrated development of the Potato Production and Marketing Sector'

From outputs to outcomes to impact

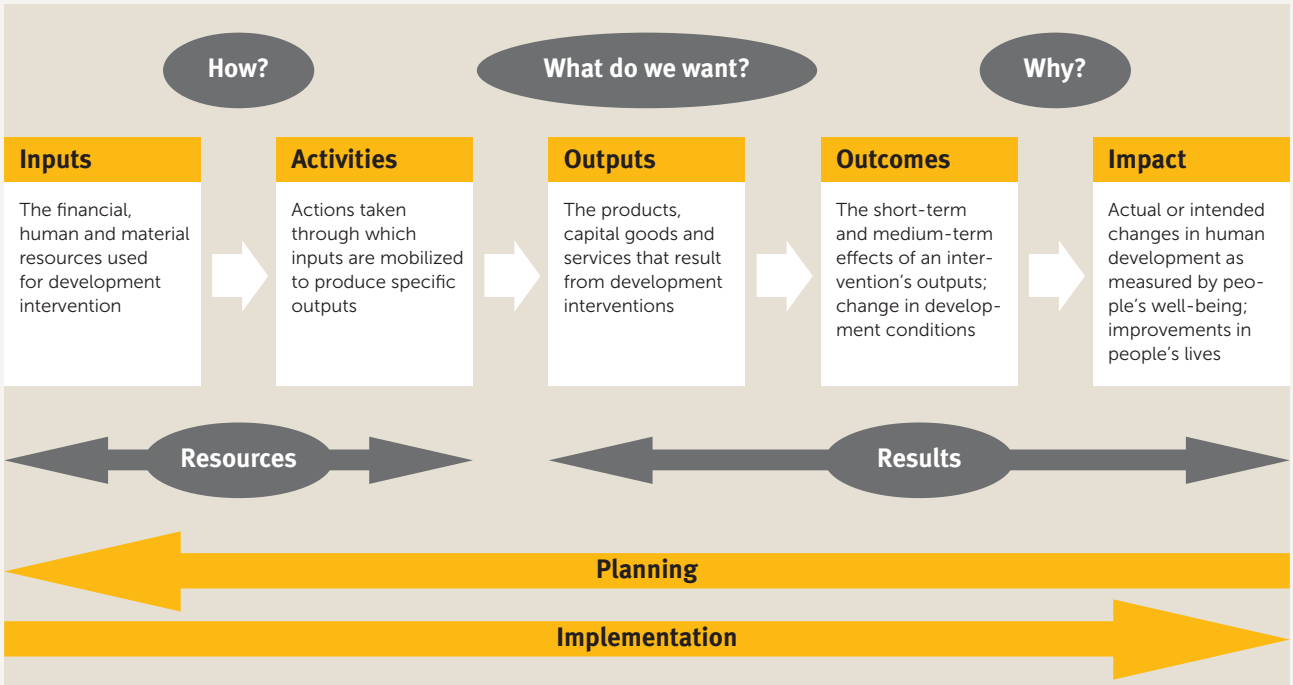
Logical models that establish indicators to measure project outputs have their place. But over time we have evolved and loosened up these models, and have really started to look a lot deeper than output level, towards outcome, and ultimately impact. 'Achieving impact' has a ring of the Holy Grail about it – everyone wants to achieve impact, even though we are not often entirely sure what it will look like until we find it.

Let's quickly move through the impact chain by way of an example: outputs could refer to the number or farmers trained. Outcomes might include how the training was valued and changes in farmer knowledge and practice. Impacts are about what this all means to farmers and their households, such as increased income, greater food security, and perhaps improved well-being and empowerment. This is obviously an oversimplified scenario (and there is often debate about what exactly qualifies as an outcome or impact), but the example helps to illustrate an important point – when you start looking beyond output, outcomes and impact often become more abstract. They are both more difficult to measure and to attribute to an intervention.

Measuring changes in well-being and empowerment, for example, require proxy indicators, since asking someone straight up if their well-being has improved does not seem very reliable and could be interpreted in all sorts of ways. Even measuring changes in income can be tricky – farming households might have several dynamic sources of income, self-reported income is well-known to be unreliable, yield estimations range with the vagaries of the weather, and price fluctuations roll with the seasons.

The most widely shared definition of impact is that of the OECD-DAC Glossary (2002), which defines it as: *'positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended'*.

This definition stresses the search for any effect, not only those that are intended, that the effects of interest are somehow caused by the intervention; that there is the possibility of all kinds of links between the intervention and effect, and that long term effects are important.



Source: UNDP Handbook on Planning, Monitoring and Evaluation for Development Results (2009, p55)

Context and the counterfactual

As practitioners and policymakers have strived towards achieving 'real' and 'sustainable' impact, evaluators have been right there with them trying to measure that impact. On first thought, one might wonder how difficult it could possibly be to show how well an intervention performed (achieved impact) and to be able to say whether or not it was the intervention that caused this change.

But think a little deeper and it quickly becomes apparent that such simple questions become minefields, for the reason that communities are socially, economically and politically complex and dynamic structures. Environmental differences between communities only add to the complexity. Put simply, development interventions do not take place in a bubble – there are many other things going on at the same time and place as the intervention, some acting as enablers and some as barriers. This makes cause and effect rather difficult to isolate.

A simple illustration can be found in the work of the PADev team, a Dutch based research team comprising academic and NGOs (see www.padev.nl). In their work in northern Ghana and Burkina Faso they found participants could recall a high number of interventions (50 to 100) by all sorts of actors (government, NGO, religious, private) around even the sleepiest villages. The point is that these may also be contributing to the same impacts as the studied intervention. To make these visible, one only needs to make a serious effort to ask local people about them. Accounting for other interventions in the same area is one issue, but a much more difficult proposition is accounting for the role played by social, economic, political, religious, ethnic and environmental dynamics.

This is important because any effort to measure impact implicitly implies the application of counterfactual logic – in other words, 'what would have happened without the intervention?' While this cannot be observed empirically (we cannot run a test in an alternative universe!), most approaches attempt to account for the counterfactual in their design. For if some or all of the change (impact) would have happened regardless, then impact cannot be fully attributed to the intervention although it might well have contributed to it (White 2009).

This means that any idea of measuring cause and effect as though it is linear is virtually impossible, because the multitude of influences is so vast and processes of change are so complex and dynamic. Causality is multi-causal and non-linear and systems are open to 'external' influences which interact with each other. Context changes everything.

Sound evaluation designs must recognize the importance of context since this can be key to drawing valid lessons for other settings. Unless we understand the role of context, the research will have weak external validity. We might know that some intervention works some place, but lack the confidence that it will work somewhere else.

Impact assessment therefore needs to deal with contemporary interventions that are often complex, multi-dimensional, indirectly delivered, multi-partnered, long-term and sustainable. What is needed then is a rigorous approach for measuring impact – but what does this mean in practice?

From vanilla evaluations to rigorous impact assessments

With an energised quest for impact, the language used by many organisations has shifted from classic vanilla evaluations to calls for 'rigorous impact assessments'. While it is easy enough to agree that an impact assessment should look at impact, not everyone agrees on how one should go about this. There is intensive debate in academic circles around what the appropriate methodologies for impact evaluation are. Practitioners too have their own strong views, even if they often have less time to write papers on the matter.

On one side of the debate are proponents of experimental and quasi-experimental research designs who argue that theirs is the rigorous approach, employing quantitative methods (usually large scale surveys). On the other side are proponents of more general methodologies, which might include some quantitative data collection, but is often much more qualitative (interviews, focus group discussions, participatory mapping etc.) (CDI, 2013, DFID 2012). Such is the vigour in the debate about rigour, William Easterly has referred to this as 'the civil war in development economics' (Easterly 2010).

There is little question that proponents of (quasi)experimental designs have held the upper hand in high level debates about what impact assessment should be, and what approaches can be deemed sufficiently rigorous to measure impact and ascertain attribution, given all of the aforementioned complexity surrounding an intervention's context. Affectionately referred to sometimes as 'randomistas', proponents of (quasi)experimental designs have asserted that their approach is the 'gold standard', originating as it does from the tradition of bio-medical research using randomised control and treatment groups (Woolcock, 2009). It is without question that this school of thought has made a significant contribution to developing the field of impact assessment.

Quasi-experimental designs are usually characterised by large sample sizes of survey data (often upwards of 1000 surveys). Proponents argue such methods are best for measuring what changed in complex contexts, due to their statistical power, but also because randomization is the only means to ensure unobservable selection bias is accounted for. Some development interventions naturally lend themselves to (quasi) experimental designs, whereas others do not. For example, changes in infant mortality rates from investments in an immunisation campaign may be relatively straightforward using (quasi)experimental methods. On the other hand, using such methods to evaluate good governance, institutional processes or human security is arguably less suitable.

While quasi-experimental designs and quantitative methods have dominated, another loose school of thought has passionately argued that impact assessment should not be defined by any particular design, method or philosophy, and that it is about selecting an appropriate method that can link causes and measured effects to explain not only the 'what', but also the 'how' and 'why'. This approach highlights the importance of theory in impact assessment. Indeed, qualitative methods have a significant role to play in understanding more about processes of change, which can then be generalised beyond a particular context (CDI, 2013; White 2009). Furthermore, qualitative data from a range of beneficiaries and stakeholder positions brings their voices into the research in terms of how an intervention was valued. It is not at all uncommon, for different beneficiaries groups to hold differing and nuanced views of an intervention. Qualitative approaches can highlight local power relations, and the reasons for regional variance in impact. Qualitative methods can also be setup to probe on what other interventions are occurring in the area, what trends have been occurring recently and their cause, and get a better picture of the context generally. Qualitative methods are also much more flexible to picking up on unintended and unplanned for consequences and changes that were not considered in the original research design.

Experimental design:

Subjects (families, schools, communities etc.) are randomly assigned to project and control groups (some randomly receive the project, others do not). Questionnaires or other data collection instruments (anthropometric measures, school performance tests, etc.) are applied to both groups before and after the project intervention. Additional observations may also be made during project implementation.

(World Bank, 2004, p.24)

Quasi-experimental design:

Where randomization [of who receives the project] is not possible, a control group is selected which matches the characteristics of the project group as closely as possible. Sometimes the types of communities from which project participants were drawn will be selected. Where projects are implemented in several phases, participants selected for subsequent phases can be used as the control for the first phase project group.

Photo: Roger Bymolt

Conducting surveys in Ethiopia using digital tablets, for the CFC project 'Wealth creation through integrated development of the Potato Production and Marketing Sector'

There is much more to say on the various strengths of each school of thought. But the point is that qualitative methods have a legitimate role to play when assessing impact, and offer tools more suited to answering the hows and whys of things than quantitative methods.

The elephant in the room

The elephant in the room is the cost of an impact assessment. Cost depends to a great extent on the methods employed and the length of time required for data collection in the field. Obviously complexity affects the costs of research design and analysis too, but let's stay with fieldwork costs here. Quasi experimental designs often call for a baseline and endline assessment, with treatment and control groups, surveying upwards of 1000 respondents so that findings can be tested as statistically significant (meaning there is a very low likelihood that measured impact was caused by chance, or another factor).

A research design for an impact assessment with this level of sophistication will likely cost upwards of 100,000 EUR, and could be considerably more depending on the design (even when using qualified local teams of enumerators and consultants for data collection) (CDI, 2013; World Bank, 2004).

Many organisations, like CFC, have picked up on the shifting semantics from evaluation to impact assessment, and have also picked up on the broad discourse around 'best practice' or 'gold standards'. This is reflected in the growth in calls for tenders that apply the language of impact assessment, and ask for baselines, endlines, control and treatment groups. They have the best of intentions to measure impact rigorously. However, budgets often don't fit the scope of their ambitions.

There is obviously a huge range of intervention types and budgets, but it is probably fair to say that the total resources allocated for most projects, programmes and impact investments is under €2 million (US$ 2.5 million), and more

Photo: Roger Bymolt

Value chain mapping in an evaluation in Transmara, Kenya

commonly under €1 million (US$ 1.25 million). There are no hard and fast rules about how much should be spent on monitoring and evaluation but typically this budget might be 2-5% of the budget. This would leave roughly €20,000-€50,000 for a €1 million project – a fair chunk of money, yet insufficient for an impact assessment employing a quasi-experimental which can withstand academic scrutiny. Sometimes a sub-set of projects under an umbrella programme can be sampled and given a higher budget, but on the whole programme managers are conscious not to misallocate scarce resources.

So what to do? Some try to take shortcuts and ask for 'rapid' impact assessments using quantitative methods. Many only conduct ex-post evaluations at the end of the project and try to use recall data as a baseline for measuring change. Alternatively, this is where a bulging toolbox of other qualitative methods comes in. Let's be clear, these are not a perfect substitute for large scale, statistically significant studies. But they can still produce evidence which is sufficiently convincing to make programming and funding decisions on, provided the impact is large enough.

Mixing methods

Recently, some of the fire in the debate over designs and methods for impact assessment has cooled just enough for evaluators from both quantitative and qualitative traditions to give a little nod of acknowledgement to each other. They are, of course, searching for the same thing – knowledge on what works and why (Garbarino & Hoolland 2009). Mixed methods can be a good approach under which they can collaborate.

Mixed methods approaches to impact evaluation are becoming more popular for several reasons. There is more demand from donors to understand both what worked and why, and so demands are being made for studies employing complementary quantitative and qualitative components. A fact often left unmentioned is that understanding processes of why an intervention worked lends additional credibility to claims of attribution (or contribution) than 'just' statistical methods (which are not infrequently based on more assumptions, incomplete datasets, or suspect quality data than researchers like to admit). Furthermore, increasing numbers of practitioners are also seeing possibilities in the way mixed methods can be applied, and are pushing for donors and clients to give them the green light to deploy these.

Mixed methods sound like a great idea, and certainly can be, but it again depends on what we are actually mixing. We should be clear that mixed methods is not one thing – in fact it can be almost anything that combines quantitative and qualitative methods. Obviously it can imply a full scale quasi experimental research design with some light qualitative focus group discussions, or the opposite – full scale qualitative research with a much smaller sample of quantitative research. So the point is that mixed methods is actually about integrating methods in a way that makes sense to the research questions being asked, the type of intervention being studied and the resources available.

When it comes to lower cost impact assessments (in the broader definition of the term), mixed methods can be one approach to collecting a modest data set of several hundred surveys, while at the same time running focus groups with well-designed exercises to elicit in-depth views across a strata of beneficiaries. This, coupled with key informant interviews, should generate sufficient data to assess those interventions with more modest budgets. More importantly, when analysis of the qualitative and quantitative datasets leads to the same conclusions (as they should), there is a kind of triangulation which supports the case made for impact and attribution. Obviously this is easy to say and much more difficult to achieve in practice, as mixed methods designs come with their own unique challenges, particularly around logistics and the capacity of researchers to be a 'jack of all trades'.

Happily, the growth in online resources is helping build sector capacity, providing more options and better support to evaluators. Well known websites such as http://mymande.org and http://betterevaluation.org/ are important knowledge and resource hubs, while organisations such as 3ie have become instrumental not just for funding impact evaluations, but for coordinating and promoting events, disseminating research and driving debate. These are just a few cherry picked examples from an energised sector.

Technology, too, has also helped. For example, digital tablets are becoming a more common means of collecting survey data, which can be pushed to a cloud server whenever there is Wi-Fi or 3G available. Open Data Kit (ODK) is one good example of a free and open-source set of tools which help organizations author, field, and manage mobile data collection solutions. Technology such as this helps to speed up the process, saving time and some costs (no transcribing from paper to digital) and eliminates errors when transcribing.

The takeaway

The takeaway messages are that evaluation has come a long way, and practitioners, policymakers and evaluators are more engaged than ever in both achieving and measuring impact. Impact assessments are important for accountability, but even more so for learning lessons that can be applied when replicating or scaling up, to build up an evidence base on good practice, and to influence policy. This naturally implies utilization of findings, if impact assessments are going to be worth the investment. A topic for another day is how evidence generated by impact assessments should be disseminated and presented, so that key findings can be easily digested without being buried in depths of heavy reports. Separate summary briefs which are clearly written with a minimum of jargon and are perhaps embellished with rich photos or infographics are much more likely to be read, and shared on social media.

The final word is that 'impact assessment' is not a term confined to a particular design or method. What is most important is to apply a design which is appropriate to the intervention, and of course, that elephant in the room, the budget.

References

Allmark, P., Boote, J., Chambers, E., Clarke, A., McDonnell, A., Thompson, A., Tod, A. (2009) *Ethical issues in the use of in-depth interviews: literature review and discussion*. Research ethics review, 5 (2), 48-54.

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

Better Evaluation (2014) *Manage an Evaluation or Evaluation System: Determine and Secure Resources*. Available from: http://betterevaluation.org/plan/manage_evaluation/determine_resources (22-07-2014)

Bamberger (2009) *Alternatives to the Conventional Counterfactual. Summary of Session 713 Think Tank*. American Evaluation Association: Orlando.

CDI (2013) Impact Evaluation: *Taking Stock and Looking Ahead. Conference Report*. Center for Development Innovation, Wageningen UR.

CGDev (2006) *When Will We Ever Learn: Improving Lives Through Impact Evaluation. Report of the Evaluaiton Gap Working Group*. Center for Global Development: Washington D.C.

DFID (2012) *Broadening the Range of Designs And Methods For Impact Evaluations*. Working Paper 38. Department for International Development: London.

Easterly, W. (2010) *What works in Development? – Thinking Big and Thinking Small*

Garbarino & Holland (2009) *Quantitative and Qualitative Methods in Impact Evaluation*

IPEN/UNICEF (2006) *New Trends in Development Evaluation*. Evaluation Working Papers Issue 5. UNICEF Regional Office for CEE/CIS and IPEN: Geneva

Kushner, S. (2005) *How does Evaluation Create Options to Enhancce Social Justice?* In: The Evaluation Exchange. Vol. XI, No. 3, Fall 2005.

Prowse, M. (2007) *Aid Effectiveness: The Role of Qualitative Research in Impact Evaluation*. Bacground Note. Overseas Development Institute: London.

Ravallion, M. (2008) *Evaluation in the Practice of Development*. Pollicy Research Working Paper 4547. The World Bank: Washington D.C.

Ton, G. (2012) *The mixing of methods: A three-step process for improving rigour in impact evaluations*. Evaluation 2012 18:5.

White (2009) *Some Reflections on Current Debates in Impact Evaluation*. Working Paper 1. International Initiative for Impact Evaluation (3ie): New Delhi.

Woolcock (2009) *Toward a Plurality of Methods in Project Evaluation: A Contextualised Approach to Understanding Impact Trajectories and Efficacy*. Journal of Development Effectiveness 1:1, 1-14.

World Bank (2004) *Monitoring & Evaluation: Some Tools, Methods and Approaches*. World Bank: Washington D.C.

World Bank (2011) *Impact Evaluation in Practice*. World Bank: Washington D.C.

Acknowledgements

This working paper was prepared thanks to the funding made available by the Common Fund for Commodities (CFC) and featured in the CFC annual report 2013. Thanks to Just Dengerink for assisting in the literature review.

KIT | Sustainable Economic Development & Gender

KIT Working Papers aims to share the results of work of KIT and its partners with development practitioners, researchers and policy makers. We welcome your feedback. Readers are encouraged to reproduce, share and quote this paper, provided due acknowledgement is given to KIT.

Correct citation: Bymolt, R. 2015. Impact assessment and the quest for the Holy Grail. KIT Working Paper 2015: 3.

© Royal Tropical Institute 2015

www.kit.nl/sed/publications