Introduction – Rethinking Impact Evaluation for Development

Barbara Befani, Chris Barnett and Elliot Stern

Abstract This *IDS Bulletin* is the first of two special issues presenting contributions from the event 'Impact Innovation and Learning: Towards a Research and Practice Agenda for the Future', organised by IDS in March 2013. The initiative, as well as these two issues, represent a 'rallying cry' for impact evaluation to rise to the challenges of a post-MDG/post-2015 development agenda. This introduction articulates first what these challenges are, and then goes on to summarise how the contributors propose to meet these challenges in terms of methodological and institutional innovation. Increasingly ambitious development goals, multiple layers of governance and lines of accountability require adequate causal inference frameworks and less ambitious expectations on the span of direct influence single interventions can achieve, as well as awareness of multiple bias types. Institutions need to be researched and become more *impact-oriented* and *learning-oriented*.

This special issue of the *IDS Bulletin* is the first of two that follow an event entitled 'Impact, Innovation and Learning: Towards a Research and Practice Agenda for the Future', held in March 2013 at the Institute of Development Studies.[1] This event brought together a distinguished group of international scholars and practitioners from academic institutions, donor country agencies, and multilaterals. It situated development evaluation in general, and impact evaluation in particular, in the specific setting of today's complex and changing international development context.

The world faces many new challenges in order to address poverty in the twenty-first century. While aid volumes from Organisation for Economic Cooperation and Development (OECD) countries are at historic highs (OECD 2014), the prospect of eliminating poverty and reducing inequality in developing countries remains stubborn. Indeed, while there has been significant progress towards the Millennium Development Goals (MDGs), there has been insufficient achievement in many areas, particularly in sub-Saharan Africa (UN 2014).

As we consider the post-2015 agenda, bilateral and multilateral aid flows now operate in a remarkably changed environment – an era of rapid change and increasing uncertainty. New actors have emerged from more traditional forms of aid cooperation, with philanthropic foundations and other private donors playing an increasingly important role. The rising powers of China, India, Brazil and elsewhere now perform an increasingly influential role in international development, while poverty persists in many Middle Income Countries (MICs) that are no longer prioritised by traditional forms of aid flow (Sumner 2010). Trade, investment and diplomacy – as well as trans-border issues like climate change, migration and terrorism – increasingly complicate the effectiveness of aid cooperation.

In this context, understanding and evaluating the impact of international development – and especially attributing effectiveness to specific interventions – is an increasingly challenging, but at the same time pressing, concern: tight budgets, greater demands for accountability, and a gradual cultural shift towards evidence-based policy have all served to reinvigorate a focus on measurement and evaluation. As a consequence, particular evaluation methods, such as experiments and quasi-experiments, have received special attention for: (a) their ability to produce findings that can be assessed according to clear quality standards, and (b) their ability to demonstrate causal links between the intervention and outcomes. The position of many donors (as

IDS Bulletin Volume 45 Number 6 November 2014
Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA

demonstrated in a series of methodological guides, such as Gertler *et al.* 2011; HM Treasury 2011 and USAID 2011) has more or less explicitly identified a hierarchy of methods, ranked by their degree of 'rigour', where rigour is broadly intended as lack of 'bias'. At the top of this hierarchy lie randomised controlled trials, followed respectively by quasi-experiments, mixed methods and qualitative methods. More or less explicitly, these rankings postulate that: (a) quantitative methods hold a superior status in comparison with qualitative methods; and (b) causal inference is exclusively the attribution of one effect to one cause, where the cause is the intervention and the effect is the 'net' or additional effect attributable to the intervention.

These positions have been somewhat heavily criticised in a series of studies and initiatives (BetterEvaluation 2014; UNEG 2013; Stern *et al.* 2012), and their limitations highlighted in the following terms: (a) lack of concern for the different tasks and purposes connected to development impact evaluation (BetterEvaluation 2014; Cartwright and Hardie 2012); (b) failure to draw on the latest academic literature for a broader range of methods and causal frameworks (Stern *et al.* 2012); (c) inability to contribute to programme improvement and transferability of lessons learned (in particular to predict programme performance in the future and in other sectors/areas/contexts) (Cartwright and Hardie 2012; Stern *et al.* 2012); (d) the limited number of programmes where experimental methods are applicable, that is, where their requirements are met; in particular, the rare times that reconstructing a plausible counterfactual is possible, or feasible given the constraints most evaluators work within; and finally, (e) failure to capture the increasingly ambitious, complex, multifaceted and long-term dimensions of development goals (Woolcock 2013; Stern *et al.* 2012).

The event in March 2013 sought to raise the level of the debate by sketching out the contours of a research and practice agenda that would meet the increasing – and sometimes contrasting – demands for evidence about development programmes and projects that work; evidence that would serve both accountability and learning purposes; that would speak to recipients as well as donors; that would capture the increasingly complex ambitions of development goals; and that would fit within the multiple governance layers and lines of accountability of the new, post-Paris declaration and post-Busan institutional settings of development assistance.

In collecting some of the contributions made to this event, this and the next special issue of the *IDS Bulletin* show that such an agenda is packed, and it goes beyond innovation on research methods. Indeed, methodological innovation is tightly linked to the new requirements of development impact evaluation, which in turn stem from systemic, historical changes, partly reflected in the reform of aid assistance.

Hence, the **first question** addressed by the contributors to this issue is: 'How does the new purpose of development cooperation change the way we approach evaluation?' What are the new goals, and how should evaluators respond?

In his article, Robert Picciotto maintains that the focus of development is now not merely on economic growth but on quality growth, equity, socially inclusive and environmentally sound strategies, and on all three dimensions of wellbeing (material, relational, perceptual). Many of these goals, including their intermediate goals, are hard to reach in the short term and even when reached are hard to measure, like empowerment. And the impact of development interventions on the achievement of these goals, potentially reached through institutional reform or cross-sector (budget) support, is even harder to assess.

Another dimension of intractability, Picciotto argues, is that in the new aid architecture, delivery is both complicated and complex, with emerging coalitions working in diverse partnership configurations. Even providing a clear and complete definition of an intervention is difficult sometimes, let alone isolating and measuring its contribution to a specific outcome.

Along similar lines, Rogers and Peersman recommend that attention be paid to 'the larger map of development', which includes 'not just donor-funded projects, but country-led programmes and policies, public–private partnership projects and civil society development interventions', and that a range of different possible users for impact evaluations, from donors and national governments to decentralised levels of government; non-governmental organisations; the private sector; and communities are kept in

mind. The authors suggest that since individual projects are increasingly designed within broader programmes; programmes within policies; and policies within strategies, instruments and tools should be in place to evaluate at all scales, beyond the single project or generic group of projects.

The methods currently enjoying the best reputation in impact evaluation are not optimised to meet these demands. In particular, they do not address the multiplicity of contributions to development outcomes, their interrelationships, or their complex trajectories over a long period of time.[2] Hence, the **second question** addressed is methodological: 'What innovation do we need in impact evaluation methods to meet the new challenges?'. Rogers and Peersman identify this as the 'practice' dimension of the agenda ('how impact evaluation is actually undertaken').

The contributors to this issue of the *IDS Bulletin* propose three directions: taking non-counterfactual causal inference seriously and shifting the focus from 'assessing impact' to 'assessing confidence' (about impact); becoming more conservative on the type of outcomes that can be credibly attributed to an intervention (immediate and intermediate rather than long-term); and spending more time and energy on becoming aware of the multiple sources of bias rather than focusing on reducing one particular bias type.

The first direction involves exploring non-counterfactual approaches to assessing causality. Here, the rigorous application of generative, mechanism-based causality is advocated by Befani and Mayne in a comparison between two generative causal inference methodologies. The authors show that non-counterfactual causal inference can be based on probability theory and stake its claim to robustness with the same strength that counterfactual inference based on statistics can achieve. The key lies in a transparent application of the Bayes' formula, coupled with the tests of process tracing, performed within the overarching evaluation approach of contribution analysis. The main message is that it should be possible and desirable to shift our focus from 'assessing impact' to 'assessing our confidence' (that the intervention had an impact).

The second direction acknowledges that in many cases it is simply impossible to rigorously attribute long-term outcomes. In such cases, Ton *et al.* suggest drawing a boundary between the sphere of direct influence of the intervention and the sphere of indirect (and harder or impossible to measure) influence. Their article makes an important theoretical contribution, proposing that impact evaluation designs include a clear discussion of where this boundary lies, and recommending that this boundary be closer to the intervention rather than the ultimate outcomes that are often referred to as 'impacts'. In particular, they recommend that impact evaluations focus on analysing and testing 'proximate' causal linkages, or those between the intervention and the immediate, or at most, intermediate outcomes, while the more distant connection with the long-term outcomes be tested within the domain of existing literature or theory. In other words, only the former analysis can be carried out empirically in the course of an evaluation, and this is where the majority of resources and energy should be allocated.

But the quality of evidence does not only depend on having a wide range of options for causal inference or on understanding what can be causally attributed and what cannot. One fundamental quality attribute to an evaluation design is the identification and acknowledgement of a broad range of bias types. On this note, Camfield *et al.* introduce us to the discovery of several different types of bias that can potentially arise in an impact evaluation, along with an explanation of how the reduction of these multiple sources of bias can be managed. Traditionally, bias has been conceived of mainly as selection bias (White 2013; Deaton 2009), and various techniques have been adopted to reduce it such as randomisation or propensity score matching. Quantitative methods have been preferred because they allow a precise estimate of the type of bias that is due to chance (aka random error), which is reduced by increasing sample size. However, as Camfield *et al.* maintain, if we conceive of quantitative impact evaluations as hypothesis tests with both Type I and Type II errors, aimed at testing whether the intervention had no impact (the null hypothesis) or rather had some (the alternative hypothesis), we can observe a systematic bias towards minimising Type II error: or the risk that the intervention actually had an impact while this is not recognised by the test (the evaluation). This is equivalent to a systematic bias in overestimating

the impact of interventions, even when applying quantitative methods. In addition to which, the article presents a long list of mostly qualitative, cognitive biases, categorised under 'empirical' (sensitivity to patterns, attribution error, self-importance, halo effect), 'researcher' (allegiance or experimenter bias, conservative bias, standpoint or positionality, similar person bias), 'methodological' (availability bias, diplomatic bias, courtesy bias, exposure bias, bias caused through multiple mediation and distance from data generation), and 'contextual' (friendship bias, pro-project bias). Some of these biases are argued to be systemic and linked to the politics of evaluation as played by institutions which are resistant to change; cognitive dissonance would be reduced in undesired ways, at times producing institutional changes that are only superficial.

These 'failures' of impact evaluation raise the issue of whether – even when the right development strategies are designed and the appropriate methods known – the current impact evaluation system has sufficient capacity to implement the required changes. Therefore, the **third question** addressed by the workshop participants is: 'Is the system fit for purpose?'. Here, the 'system' (or what Rogers and Peersman call the 'enabling environment') refers to the institutional settings within which impact evaluation and development evaluation processes take place: in the words of Rogers and Peersman, the 'policies, guidelines, guidance, formal and informal requirements and resources'.

This systemic aspect is explored through several lenses, both in this and in the forthcoming issue of the *IDS Bulletin*. (The latter includes those contributions which tackle the topic using a specific language from the systems thinking tradition, while the articles included here address a more general audience.) Firstly, Vaessen *et al.* provide an overview of the challenges of the current system from the perspective of commissioners of United Nations (UN) evaluations. They argue that monitoring and evaluation functions of UN organisations need to become more *impact-oriented*. This is both to provide evidence on the performance of the organisation (and progress towards impact), as well as to allow for evaluation units to design and conduct impact evaluations based on solid data and on those parts of the portfolio where they are most needed and useful. The authors identify three solutions: improving the quality of impact-related evidence at activity and project level; strengthening the causal logic underlying interventions; and strengthening the aggregation and synthesis of evidence. Some of the specific suggestions, such as using a theory-based review approach combined with a standardised rating system to develop insights about impact at portfolio level; improving the causal logic at higher levels of intervention using nested theories of change; and developing and using analytical tools to aggregate/synthesise patterns of impact at higher levels of intervention, show that becoming *impact-oriented* will require becoming *learning-oriented*.

Similarly based on the author's personal experience with commissioning evaluations, Ole Winckler Andersen advocates for more rigorous analytic work and empirical evidence of the evaluation commissioning and execution processes. This analytic work would ideally follow a political economy perspective which explains the behaviour of evaluators and commissioners on the basis of their preferences and strategies, as well as the resources available to them (normative, legal, economic and financial). The article argues that only by knowing more about evaluation processes, described in terms of interaction between evaluators and commissioners taking place within a context of opportunities and constraints, will we be able to explain why many evaluations are of inadequate quality, and thus improve the system so that it becomes more 'fit for purpose'. In this sense the article opens up an almost entirely new avenue of research (duly noted in Rogers and Peersman, this *IDS Bulletin*), while contrasting the popular idea that evaluation quality is ultimately dependent on the characteristics 'of the evaluator' as isolated from the context that s/he operates in.

Along these lines, Camfield *et al.* acknowledge some form of 'failure' of 'evaluation implementation', and go to greater length in analysing it, using the notion of 'isomorphic mimicry' to explain the general trend of public agencies emulating private sector (for profit) organisations. Beyond research on evaluation processes and methodological solutions like transparency and reflexivity, proposed institutional solutions include peer review and ethical codes. In addition, Rogers and Peersman suggest that the research agenda should include not just the enabling environment for conducting

impact evaluations, their practice and products, but also the impacts of impact evaluation processes on users and, more generally, the uses of impact evaluation products.

Given the experimental nature of innovation efforts, some of the ideas presented in these contributions might be closer to 'proof-of-concept' studies than fully-baked academic products. Nonetheless, given the pressing concerns that move them, we deem it not only worthwhile, but also necessary to expose these ideas to the community, hoping to strengthen the debate in a way that will ultimately lead to more consolidated guidance.

In summary, this *IDS Bulletin* presents a 'rallying cry' for impact evaluation to rise to the challenges of a post-MDG/post-2015 world. The past decade has seen a resurgent interest in addressing pressing 'impact questions' about the effectiveness of development assistance, and yet while these questions are as relevant today as they ever were, the context is changing: increasingly ambitious development goals, multiple layers of governance and lines of accountability, emergent and increasingly influential actors, and a changing way in which interventions are undertaken.

Those methods currently enjoying the best reputation are not necessarily optimised to address the multiplicity of development outcomes, their interrelationships, or the complex pathways towards long-term impact. This is fertile ground for a new research and practice agenda: one that can better enable impact evaluation to meet the new purposes of development cooperation; one that can innovate around methodological designs and practice to address increasingly complex challenges; and one that will help us better understand and improve evaluation systems. Ultimately though, the success of such an emerging agenda rests on whether we can make better use of evaluative evidence to have a real impact on the lives of the poorest and most marginalised.

Notes

1 The second issue will tackle the application of systems thinking and complexity science to impact evaluation and learning, and will be published in January 2015.

2 The January 2015 issue of the *IDS Bulletin* is specifically focused on the complex and systemic dimensions of these challenges.

References

BetterEvaluation (2014) *Rainbow Framework Overview*, www.betterevaluation.org/plan (accessed 4 September 2014)

Cartwright, N. and Hardie, J. (2012) *Evidence-Based Policy: A Practical Guide to Doing It Better*, Oxford: Oxford University Press

Deaton, A.S. (2009) *Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development*, Working Paper 14690, Cambridge MA: National Bureau of Economic Research

Gertler, P.J.; Martinez, S.; Premand, P.; Rawlings, L.B. and Vermeersch, C.M.J. (2011) *Impact Evaluation in Practice*, Washington DC: World Bank

HM Treasury (2011) *The Magenta Book: Guidance for Evaluation*, London: HM Treasury

OECD (2014) *Aid to Developing Countries Rebounds in 2013 to Reach an All-time High*, www.oecd.org/newsroom/aid-to-developing-countries-rebounds-in-2013-to-reach-an-all-time-high.htm (accessed 4 September 2014)

Stern, E.; Stame, N.; Mayne, J.; Forss, K.; Davies, R. and Befani, B. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*, DFID Working Paper 38, London: Department for International Development

Sumner, A. (2010) *Global Poverty and the New Bottom Billion: Three-quarters of the World's Poor Live in Middle-income Countries*, IDS Working Paper 349, Brighton: IDS

UN (2014) *The Millennium Development Goals Report 2014*, New York NY: United Nations

UNEG (2013) *Impact Evaluation in UN Agency Evaluation Systems: Guidance on Selection, Planning and Management*, New York NY: United Nations Evaluation Group

USAID (2011) *USAID Evaluation Policy: Evaluation, Learning from Experience*, Washington DC: United States Agency for International Development

White, H. (2013) 'An Introduction to the Use of Randomised Control Trials to Evaluate Development Interventions', *Journal of Development Effectiveness* 5.1: 30–49

Woolcock, M. (2013) 'Using Case Studies to Explore the External Validity of "Complex" Development Interventions', *Evaluation* 19.3: 229–48