Impact Evaluation Notes No. 1. March 2012

INTRODUCTION TO IMPACT EVALUATION

Patricia J. Rogers, RMIT University (Australia) and BetterEvaluation

This is the first guidance note in a four-part series of notes related to impact evaluation developed by InterAction with financial support from the Rockefeller Foundation. The other notes in this series are: Linking Monitoring & Evaluation to Impact Evaluation; Introduction to Mixed Methods in Impact Evaluation; and Use of Impact Evaluation Results. The complete series can be found on InterAction's website at: http://www.interaction.org/impact-evaluation-notes.

InterAction
A UNITED VOICE FOR GLOBAL CHANGE

THE ROCKEFELLER FOUNDATION

ACKNOWLEDGEMENTS

This guidance note has benefited from feedback from members of InterAction's Evaluation and Program Effectiveness Working Group. Information about evaluation methods in Guidance Note 1 has been drawn from BetterEvaluation, an international collaboration for sharing information to improve evaluation. More information about each of the methods can be found on the website www.betterevaluation.org.

Photo: David Darg

Contents

Introduction

Credible and appropriate impact evaluation can greatly improve the effectiveness of development. The increasing emphasis on impact evaluation in development has led to many questions. What constitutes credible and appropriate impact evaluation? How should impact evaluations be managed? What measures and data sources are appropriate? How can qualitative and quantitative data be effectively combined in impact evaluation? What should be done to support the appropriate use of impact evaluations? What are the implications of the increasing focus on impact evaluation for other types of monitoring and evaluation (M&E)?

InterAction has produced a series of guidance notes addressing these questions to support management, program and M&E staff in international NGOs to plan, design, manage, conduct and use impact evaluations. These notes can also inform their discussions with external evaluators, partners and funders.

This first guidance note, *Introduction to Impact Evaluation*, provides an overview of impact evaluation, explaining how impact evaluation differs from – and complements – other types of evaluation, why impact evaluation should be done, when and by whom. It describes different methods, approaches and designs that can be used for the different aspects of impact evaluation: clarifying values for the evaluation, developing a theory of how the intervention is understood to work, measuring or describing impacts and other important variables, explaining why impacts have occurred, synthesizing results, and reporting and supporting use. The note discusses what is considered good impact evaluation – evaluation that achieves a balance between the competing imperatives of being useful, rigorous, ethical and practical – and how to achieve this. Footnotes throughout the document contain references for further reading in specific areas.

1. What Do We Mean by "Impact Evaluation"?

Impact evaluation investigates the changes brought about by an intervention. Impact evaluation can be undertaken on interventions at any scale: a small, local HIV-AIDS project; an entire civil society strengthening program of an NGO; a sequence of natural resource management projects undertaken in a geographic area; or a collection of concurrent activities by different organizations aimed at improving a community's capacity.

The expected results of an intervention are an important part of an impact evaluation, but it is important to also investigate unexpected results. In this guidance note, impacts are defined as:

> the positive and negative, intended and unintended, direct and indirect, primary and secondary effects produced by an intervention. (OECD Development Assistance Committee definition)[1]

Impacts are usually understood to occur later than, and as a result of, intermediate outcomes. For example, achieving the intermediate outcomes of improved access to land and increased levels of participation in community decision-making might occur before, and contribute to, the intended final impact of improved health and well-being for women. The distinction between outcomes and impacts can be relative, and depends on the stated objectives of an intervention.

In practice, it is often helpful for an evaluation to include both outcomes and impacts. This allows earlier indication of whether or not an intervention is working – and if it is not working, helps to identify where, and perhaps why.

In this guidance note, an impact evaluation includes any evaluation that systematically and empirically investigates the impacts produced by an intervention. Some individuals and organizations use a narrower definition of impact evaluation, and only include evaluations containing a counterfactual of some kind (an estimate of what would have happened if the intervention had not occurred) or a particular sort of counterfactual (for example, comparisons with a group who did not receive the intervention). USAID, for example, uses the following definition: "Impact evaluations measure the change in a development outcome that is attributable to a defined intervention; impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change." These different definitions are important when deciding what methods or research designs will be considered credible by the intended users of the evaluation or by partners or funders.

Impact evaluation is, of course, not the only type of evaluation that supports effective development. It is important to ensure that investments in impact evaluation (in terms of time and money) are not made at the expense of monitoring or other types of evaluation – such as needs assessment, process evaluation and cost-benefit evaluation – that are also needed to inform decisions about practice and policy. Guidance Note 2 discusses how impact evaluation and these other types of monitoring and evaluation can be done in ways that support each other. For example, monitoring data can

[1] Impacts are sometimes defined quite differently. For example, the W.K.Kellogg Foundation Logic Model Development Guide (www.wkkf.org/knowledge-center/resources/2006/02/WK-Kellogg-Foundation-Logic-Model-Development-Guide.aspx) distinguishes impact in terms of its spread beyond those immediately involved in the program. Specific changes in program participants' behavior, knowledge, skills, status and level of functioning are referred to as "outcomes," and only changes to organizations, communities or systems as a result of program activities within seven to 10 years are described as "impacts."

provide a good foundation for impact evaluation, and an impact evaluation can guide the development of monitoring systems. Impact evaluation provides necessary information for cost-benefit and cost-effectiveness evaluations.

2. Why Should We Do Impact Evaluation?

The best way to undertake a particular impact evaluation depends in part on its purpose and who its primary intended users are. Some common reasons for doing impact evaluation include:

- **To decide whether to fund an intervention** – "ex-ante evaluation" is conducted *before an intervention is implemented*, to estimate its likely impacts and inform funding decisions.
- **To decide whether or not to continue or expand an intervention.**
- **To learn how to replicate or scale up a pilot.**
- **To learn how to successfully adapt a successful intervention to suit another context.**
- **To reassure funders, including donors and taxpayers (upward accountability), that money is being wisely invested** – including that the organization is learning what does and doesn't work, and is using this information to improve future implementation and investment decisions.
- **To inform intended beneficiaries and communities (downward accountability) about whether or not, and in what ways, a program is benefiting the community.**

Guidance Note 4 discusses in more detail how to support these different ways of using impact evaluation.

3. What Questions Does Impact Evaluation Seek to Answer?

An impact evaluation should focus on a small number (five to seven) of specific key evaluation questions. These are the high-level questions that an evaluation addresses, not specific questions that might be asked in an interview or a questionnaire. It is better to focus on a small number of questions directly related to the purpose than to spread evaluation resources, and users' focus, across a large number of questions. (*See box on p. 4 for examples of key evaluation questions for impact evaluation.*)

4. Who Should Conduct Impact Evaluation?

Impact evaluation can be undertaken by: an external evaluator or evaluation team; an internal but separate unit of the implementing organization; those involved in an intervention (including community members); or a combined team of internal and external evaluators.

An external evaluator can bring a range of expertise and experience that might not be available within the organization, and may have more independence and credibility than an internal evaluator. For example, the USAID Evaluation Policy sets out an expectation that most evaluations will be done by an external evaluator.

However, for some stakeholders, external evaluators are not always perceived as unbiased, as their data gathering and interpretations may be affected by their lack of familiarity with the context. In some cases, involving program stakeholders and/or

Examples of key evaluation questions for impact evaluation

Overall impact

- Did it work? Did [the intervention] produce [the intended impacts] in the short, medium and long term?
- For whom, in what ways and in what circumstances did [the intervention] work?
- What unintended impacts (positive and negative) did [the intervention] produce?

Nature of impacts and their distribution

- Are impacts likely to be sustainable?
- Did these impacts reach all intended beneficiaries?

Influence of other factors on the impacts

- How did [the intervention] work in conjunction with other interventions, programs or services to achieve outcomes?
- What helped or hindered [the intervention] to achieve these impacts?

How it works

- How did [the intervention] contribute to [intended impacts]?
- What were the particular features of [the intervention] that made a difference?
- What variations were there in implementation?
- What has been the quality of implementation in different sites?
- To what extent are differences in impact explained by variations in implementation?

Match of intended impacts to needs

To what extent did the impacts match the needs of the intended beneficiaries?

community members in conducting an evaluation can add rigor and credibility by supporting better access to data (especially key informants) and more appropriate interpretation of the data.

Three practices in particular can often produce the best quality evaluation: establishing a team of evaluators with external and internal perspectives; ensuring transparency in terms of what data are being used and how in the evaluation; and triangulation – using multiple sources of evidence (which have complementary strengths) and multiple perspectives in analysis and interpretation. It is especially useful to include local evaluation experts on the team who know the context, history and comparative interventions by other agencies.

An evaluation can be managed by an internal group (perhaps an internal steering committee, informed by an advisory group with external membership) or by a combined group. Participatory approaches to managing evaluations typically involve program staff, community members and development partners. They participate not only in collecting data, but also in negotiating the purpose of the impact evaluation, developing the key evaluation questions, designing an evaluation to answer them and following through on the results.[2]

[2] Additional sources on participatory methods include: Marisol Estrella et al. (eds), *Learning from Change: Issues and Experiences in Participatory Monitoring and Evaluation* (Brighton: Institute of Development Studies, 2000), http://www.idrc.ca/EN/Resources/Publications/Pages/IDRCBookDetails.aspx?PublicationID=348; Andrew Catley et al., "Participatory Impact Assessment," Feinstein International Center, Tufts University: October 2008, http://sites.tufts.edu/feinstein/2008/participatory-impact-assessment; Robert Chambers,"Who Counts? The Quiet Revolution of Participation and Numbers," Working Paper No. 296 (December 2007), Brighton: Institute of Development Studies, http://www.ids.ac.uk/files/Wp296.pdf.

5. How Should We Choose Methods for Impact Evaluation?

There has been considerable debate in development evaluation, and more broadly, about which methods are best for impact evaluation. These discussions reflect different views on what constitutes credible, rigorous and useful evidence, and who ought to be involved in conducting and controlling evaluations.

Some organizations and evaluators have argued that particular methods or research designs should be used wherever possible – for example, randomized controlled trials or participatory methods. Others have argued for situational appropriateness. This means choosing methods that suit the purpose of the evaluation, the types of evaluation questions being asked, the availability of resources, and the nature of the intervention – in particular whether it is standardized or adaptive, and whether interventions work pretty much the same everywhere and for everyone or are greatly affected by context.

When choosing methods, it is important to address each of six different aspects of an impact evaluation:

- **Clarifying the values** that will underpin the evaluation – what will be considered desirable and undesirable processes, impacts and distribution of costs and benefits?
- **Developing and/or testing a theory** of how the intervention is supposed to work – these are sometimes referred to as theories of change, logic models or program theory.
- **Measuring or describing** these impacts and other relevant variables, including processes and context.
- **Explaining** whether the intervention was the cause of observed impacts.
- **Synthesizing** evidence into an overall evaluative judgment.
- **Reporting findings and supporting their use.**

This guidance note discusses each of these aspects and provides information about a range of methods that can be used for them. Links to additional sources of information are provided. Guidance Note 3 discusses how a mixed method approach, combining quantitative and qualitative data in complementary ways, can be both measurement/description and explanation.

6. Clarifying Values for an Impact Evaluation

The first step is to clarify the values that will underpin the evaluation. Impact evaluation draws conclusions about the degree of success (or failure) of an intervention, so it is important to clarify what success looks like in terms of:

- **Achieving desirable impacts and avoiding (or at least minimizing) negative impacts.** For example, will the success of a road development project be judged in terms of increased access to markets, or improved access to maternity hospitals? What level of loss of habitat and biodiversity would be considered a reasonable cost for the road? What level would not be an acceptable trade-off?
- **Achieving a desirable distribution of benefits.** For example, should we judge success in terms of the average educational outcome, improvements for the most disadvantaged, or bringing a vulnerable or disadvantaged group (like young girls) up to the same level as their more advantaged counterparts?

Formal stated goals (including the Millennium Development Goals) and organizational policies are an important start to clarifying values, but are by themselves usually not sufficient. Different stakeholders may well have different views about which values should be used in an evaluation.

Some methods for clarifying the values for an impact evaluation:

Methods that help people articulate tacit values

Appreciative inquiry – key stakeholders (including program staff) recall times when the program worked particularly well, then identify the values it exemplified during those times (Using Appreciative Inquiry in Evaluation Practice).

Community surveys – individuals in the community either nominate or rate the issues that they see as most important to address.

Most significant change – a structured process for generating and selecting stories of change that identify what different individuals and groups see as the most important outcomes or impacts. (Most Significant Change)

Methods that help negotiate between different sets of values

Delphi – process that works through a series of written interactions without face-to-face context, where key stakeholders provide their opinions about what they see as important, then respond to the aggregated results (Delphi Method | Delphi Method: Techniques and Applications | Delphi Survey - Europa).

Sticky dot voting – in a face to face meeting, individuals allocate their multiple "votes" (in the form of sticky dots) across options (NRCOI Quick Tip).

7. Developing a Theory or Model of How the Intervention is Supposed to Work

It is often helpful to base an impact evaluation on a theory or model of how the intervention is understood to produce its intended impacts. This might be called a program theory, a theory of change (ToC), a results chain or a logic model. It is best to develop the theory of change as part of planning an intervention, and then to review it and revise it as necessary while planning an impact evaluation. If this has not been done by the time the intervention starts, it is possible to retroactively develop an agreed theory of change.

Depending on when the theory of change is developed, it can draw on a combination of sources: official documents and stated objectives; research into similar interventions; observations of the intervention or similar interventions; or asking different stakeholders (including planners, staff and intended beneficiaries) how they think it works (or should work).

There can be multiple theories of change – different theories showing how the intervention works at different stages, in different contexts (acknowledging effects of external influences) and for different impacts; and different theories that are developed over time as better understanding develops.

Theories of change can improve impact evaluation by helping to:

- Identify intermediate outcomes or impacts that can be observed within the time frame of the evaluation, and that are precursors to the longer-term impacts that the intervention is intended to produce.

- Identify, if an intervention was unsuccessful, where in the process it stopped working or broke down.
- Distinguish between *implementation* failure (where impacts have not been achieved because the intervention has not been properly implemented) and *theory* failure (where the intervention does not lead to desired impact even when implemented well).
- Identify what aspects of the intervention make it work, and are therefore critical and need to be continued when an intervention is adapted for other settings.
- Identify important behavioral and contextual variables that should be addressed in data collection, analysis and reporting to understand variations in impacts.
- Provide a conceptual framework for bringing together diverse evidence about a program involving a large number of diverse interventions.

Some methods for representing a theory of change:

Logical framework approach (logframe) – the classic format used in many development organizations, which uses a 4x4 matrix. The four rows are *activities*, *outputs*, *purpose* and *goal*, and the four columns are *a narrative description*, *objectively verifiable indicators (OVIs)*, *means of verification (MoV)* and *assumptions*. (The Logical Framework Approach | Logical Framework Analysis | Beyond Logframe: Critique, Variations and Alternatives)

Results chain – the intervention is represented as a series of boxes in a sequence: inputs, activities, outputs, short-term outcomes, longer-term outcomes and impacts. (Results Chain: Enhancing Program Performance with Logic Models Guide |W.K. Kellogg Foundation Logic Model Development Guide)

Outcomes chain/outcomes hierarchy/theory of change – the theory is represented as a series of intermediate outcomes leading to the final intended impacts. This format focuses attention on how change comes about, and is helpful for representing programs where different activities occur along the causal chain, not just up front. (Theory of change and logic model: Telling them apart)

Outcome mapping – this focuses on identifying the "boundary partners" – organizations or groups whose actions are beyond the control of the intervention, but are essential for the impact to be achieved – and then articulating what these partners need to do and how the intervention can seek to influence them. (Outcome Mapping | Outcome Mapping: ILAC Brief 7)

Other useful resources for developing a theory of change can be found at Developing a Logic Model or Theory of Change.[3]

A theory of change can also be used to manage potential negative impacts, or to plan an impact evaluation that measures them.

[3] The Community Toolbox, "Developing a Logic Model or Theory of Change," http://ctb.ku.edu/en/tablecontents/sub_section_main_1877.aspx.

For example, a program meant to improve agricultural productivity by encouraging farmers to apply fertilizer to their fields might lead to increased phosphate runoff and environmental damage to waterways. A balanced impact evaluation will investigate this possible impact in addition to the intended impact of improved productivity. A theory of change can be constructed to examine how an intervention might produce negative impacts. This can be used to adapt the intervention in order to minimize or avoid such negative impacts, to develop early warning indicators for monitoring purposes, and to ensure that these are included in the impact evaluation plan.

8. Measuring or Describing Impacts (and other Important Variables)

An impact evaluation needs credible evidence, and not only about impacts. Good information is also needed about how well an intervention has been implemented in order to distinguish between implementation failure and theory failure. Information is also needed about the context to understand if an intervention only works in particular situations.

It is useful to identify any data already available about impacts, implementation and context from existing sources, such as official statistics, program documentation, Geographic Information Systems (GIS), and previous evaluation and research projects. Additional data can be gathered to fill in gaps or improve the quality of existing data using methods such as interviews (individual and group; structured, semi-structured or unstructured), questionnaires (including web-based questionnaires and collecting data by cell phone), observation (structured, semi-structured or unstructured) and direct measurement (for example, of water quality against an international standard).

Descriptions of impacts should not only report the average, but also how varied the results were, and in particular report on patterns. Howard White discusses the importance of looking at heterogeneity in his 2010 article:

> *A study which presents a single impact estimate (the average treatment effect[4]) is likely to be of less use to policy makers than one examining in which context interventions are more effective, which target groups benefit most, and what environmental settings are useful or detrimental to achieving impact. Hence it could be shown that an educational intervention, such as flip charts, works but only if teachers have a certain level of education themselves, or only if the school is already well equipped with reading materials, or the pupils' parents are themselves educated.*[5]

Some sources for measures and indicators in particular sectors include:

> Catalog of survey questionnaires: International Household Survey Network – over 2,000 questionnaires that can be searched by country, date and survey types.
>
> Democratic governance – UNDP Oslo Governance Centre.

[4] The average treatment effect is an estimate of the average difference an intervention makes. For example, students in the program stayed in school an average of 2.5 years longer (compared to the control group).

[5] Howard White, "A Contribution to Current Debates in Impact Evaluation," *Evaluation* (April 2010, vol. 16 no. 2): 160.

Human Poverty Index – three indicators that relate to survival, knowledge and economic provisioning UNDP.

Millennium Development Goals – 48 technical indicators and 18 targets for the 8 goals.

Sustainable Development – UN Commission on Sustainable Development 130 indicators of social, economic, environmental and institutional aspects of sustainable development.

World Development Indicators – The World Bank has data on more than 200 countries in terms of more than 1000 indicators.

Guidance Note 3 provides more detail on specific methods for measuring or describing impacts, the use of mixed methods (quantitative and qualitative data used in complementary ways), and ways of addressing challenges in measuring or describing impacts.

9. Explaining to What Extent Observed Results Have Been Produced by the Intervention

One of the important features of an impact evaluation is that it does not just gather evidence that impacts have occurred, but tries to understand the intervention's role in producing them. It is rarely the case that an intervention is the sole cause of changes. Usually, an intervention works in combination with other programs, a favorable context or other factors. Often a group collaborates to produce a joint impact, such as when international NGOs partner with local governments and communities. Therefore, "causal attribution" does not usually refer to total attribution (that is, the intervention was the only cause), but to partial attribution or to analyzing the intervention's contribution. This is sometimes referred to as "plausible contributions."

For example, in agricultural research, impacts in terms of improved productivity can be due to a long chain of basic and applied research, product development and communication. An investment in any one of these might reasonably claim that it was essential in producing the impacts, but would not have been able to do so without the other interventions. In other words, it could have been a necessary intervention but not sufficient to bring about that impact by itself.

It can be helpful to investigate causal attribution or plausible contribution in terms of three components. The starting point is the factual – to compare the actual results to those expected if the theory of change were true. When, where and for whom did the impacts occur? Are these results consistent with the theory that the intervention caused or contributed to the results? The second component is the counterfactual – an estimate of what would have happened in the absence of the intervention. The third component is to investigate and rule out alternative explanations. In some cases, it will be possible to include all three components in an impact evaluation. In complex situations, it might not be possible to estimate a counterfactual, and causal analysis will need to depend on the other components.

Possible methods for examining the factual (the extent to which actual results match what was expected):

Comparative case studies – did the intervention produce results only in cases when the other necessary elements were in place?

Dose-response – were there better outcomes for participants who received more of the intervention (for example, attended more of the workshops or received more support)?

Beneficiary/expert attribution – did participants/key informants believe the intervention had made a difference, and could they provide a plausible explanation of why this was the case?

Predictions – did those participants or sites predicted to achieve the best impacts (because of the quality of implementation and/or favorable context) do so? How can anomalies be explained?

Temporality – did the impacts occur at a time consistent with the theory of change – not before the intervention was implemented?

Possible methods for examining the counterfactual (an estimate of what would have happened in the absence of the intervention) include:

Difference-in-difference – The before-and-after difference for the group receiving the intervention (where they have not been randomly assigned) is compared to the before-after difference for those who did not. (Difference-in-Differences)

Logically constructed counterfactual – In some cases it is credible to use the baseline as an estimate of the counterfactual. For example, where a water pump has been installed, it might be reasonable to measure the impact by comparing time spent getting water from a distant pump before and after the intervention, as there is no credible reason that the time taken would have decreased without the intervention (White, 2007). Process tracing can support this analysis at each step of the theory of change. (Process Tracing in Case Study Research)

Matched comparisons – Participants (individuals, organizations or communities) are each matched with a nonparticipant on variables that are thought to be relevant. It can be difficult to adequately match on all relevant criteria. (Techniques for improving constructed matched comparison group impact/outcome evaluation designs)

Multiple baselines or rolling baselines – The implementation of an intervention is staggered across time and intervention populations. Analysis looks for a repeated pattern in each community of a change in the measured outcome after the intervention is implemented, along with an absence of substantial fluctuations in the data at other time points. It is increasingly used for population-level health interventions. (The Multiple Baseline Design for Evaluating Population-Based Research)

Propensity scores – this technique statistically creates comparable groups based on an analysis of the factors that influenced people's propensity to participate in the program – it is particularly useful when participation is voluntary (for example, watching a television show with health promotion messages). (Propensity Scores: What, How, Why | A Practical Guide to Propensity Score Models)

Randomized controlled trial (RCT) – Potential participants (or communities, or households) are randomly assigned to receive the intervention or be in a control group (either no intervention or the usual intervention) and the average results of the different groups are compared. (Using Randomization in Developmental Economics Research)

Regression discontinuity – Where an intervention is only available to participants above or below a particular cutoff point (for example, income), this approach compares outcomes of individuals just below the cutoff point with those just above the cutoff point. (Impact Evaluation: Regression Discontinuity)

Statistically created counterfactual – A statistical model, such as a regression analysis, is used to develop an estimate of what would have happened in the absence of an intervention. This can be used when the intervention is already at scale – for example, an impact evaluation of the privatization of national water supply services.

Developing a credible counterfactual can be difficult in practice. It is often difficult to match individuals or communities on the variables that really make a difference. Randomized controlled trials can randomly create nonequivalent groups. Other methods depend on various assumptions which might not be met. In situations of rapid and unpredictable change, it might not be possible to construct a credible counterfactual. It might be possible to build a strong, empirical case that an intervention produced certain impacts, but not to be sure about what would have happened if the intervention had not been implemented. For example, it might be possible to show that the development of community infrastructure for raising fish to be consumed or sold was directly due to a local project, without being able to confidently state that this would not have happened in the absence of the project (perhaps through an alternative project being implemented by another organization). What an impact evaluation can focus on is the other two elements of causal analysis – the *factual* and *ruling out* alternatives.

The third component of understanding causal linkages is to investigate and rule out alternative explanations. Apparent impacts (or lack thereof) might reflect methodological issues such as selection bias (where participants are systematically different from nonparticipants), and contamination effects (where nonparticipants benefit from the intervention as well, reducing the difference between them and participants in terms of impacts). They might reflect the influence of other factors, including other interventions or population movements between areas assigned to receive an intervention and those without one.

Possible methods for identifying and ruling out alternative possible explanations include:

General elimination methodology – possible alternative explanations are identified and then investigated to see if they can be ruled out. (Can We Infer Causation from Cross-Sectional Data?)

Searching for disconfirming evidence/ Following up exceptions[6]

[6] Further reading: Matthew B. Miles and A. Michael Huberman, *Qualitative Data Analysis: An Expanded Sourcebook. 2nd Edition* (Thousand Oaks, California: Sage Publications, 1994).

An evaluation of the impact of legislation for compulsory bicycle helmets found that there had been a significant decline in the number of head injuries among cyclists. While this was consistent with the theory of change, an alternative explanation was that the overall level of injuries had declined due to increased building of bicycle lanes during the same period. Examination of serious injuries showed that, while the level of head injuries had declined in this period, the number of other types of injuries had remained stable, supporting the theory that it was the helmets that had produced the change. (Walter et al., 2011)

Some approaches that combine these different elements of explanation include**:**

Multiple lines and levels of evidence (MLLE) – a wide range of evidence from different sources is reviewed by a panel of credible experts spanning a range of relevant disciplines. The panel identifies consistency with the theory of change while also identifying and explaining exceptions. MLLE reviews the evidence for a causal relationship between an intervention and observed impacts in terms of its strength, consistency, specificity, temporality, coherence with other accepted evidence, plausibility, and analogy with similar interventions.[7]

Contribution analysis – a systematic approach that involves developing a theory of change, mapping existing data, identifying challenges to the theory – including gaps in evidence and contested causal links – and iteratively collecting additional evidence to address these. Guidance Note 2 provides some additional information on contribution analysis. (Contribution Analysis: ILAC Guide Brief 16 | Contribution Analysis)

Collaborative outcomes reporting – this new approach combines contribution analysis and MLLE. It maps existing data against the theory of change and fills in important gaps in the evidence through targeted additional data collection. Then a combination of expert review and community consultation is used to check the evidence's credibility regarding what impacts have occurred and the extent to which these can be realistically attributed to the intervention. (Collaborative Outcomes Reporting Technique)

An evaluation of a cross-government executive development program's impact could not use a randomized control group, because randomly assigning people to be in a control group – or even participate in the program – was impossible. Neither could the evaluation use a comparison group, because the nature of the program was such that those accepted into it were systematically different to those who were not. Instead, the evaluation used other strategies for causal explanation, including attribution by beneficiaries, temporality and specificity (changes were in the specific areas addressed by the program). (Davidson, 2006)

[7] Further reading: Patricia Rogers, "Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation", *Journal of Development Effectiveness*, 1 (2009): 217-226. Working paper version available at: http://www.3ieimpact.org/admin/pdfs_papers/50.pdf.

10. Synthesizing Evidence

It is rare to base the overall evaluative judgment of an intervention on a single performance measure. It usually requires synthesizing evidence about performance across different dimensions.

A common way to do this is to develop a weighted scale, where a percentage of the overall performance rating is based on each dimension. However, a numeric weighted scale often has problems, including arbitrary weights and lack of attention to essential elements. (The Synthesis Problem)

An alternative is to develop an agreed global assessment scale (or rubric) with intended users that can then be used to synthesize evidence transparently. The scale includes a label for each point (for example, "unsuccessful," "somewhat successful," "very successful") and a description of what each of these looks like. (The Rubric Revolution)

11. Reporting Findings and Supporting Use

The format of the evaluation report should be agreed on when the impact evaluation is being planned. Some organizations have standard report formats, including length requirements, that must be followed. In other cases, it is important to agree on a "skeleton report" of headings and subheadings well before the report is written.

Impact evaluation reports are most accessible when they are organized around the key evaluation questions, rather than reporting separately on the data from different components of data collection.[8]

The quality of evaluation reports can be enhanced by appropriate stakeholder involvement. Even where an evaluation is being undertaken by an independent external evaluator, stakeholders can be involved by providing formal responses to findings and commenting on the data and how they have been interpreted.

Where recommendations are included in evaluation reports, they need to be supported by evidence from the evaluation findings and about the feasibility and appropriateness of the recommendations. Involving relevant stakeholders in developing the recommendations can not only improve the recommendations' feasibility, but can also lead the stakeholders to both own and commit to implementing them.

The use of impact evaluation reports can be enhanced by creative reporting formats, verbal presentations, opportunities to engage with others discussing the implications of impact evaluations, and by ensuring the reports remain accessible to potential users.

Guidance Note 4 addresses the need to communicate the findings well to intended audiences.

12. When Should an Impact Evaluation Be Done?

Impact evaluations should be undertaken when there is both a clear need and intent to use the findings. If all interventions were required to have an impact evaluation, evaluators would risk either requiring an excess of resources, or spreading those resources so thin as to make evaluations superficial. A more effective strategy is to focus impact evaluation resources on interventions where they are likely to be most useful:

[8] E. Jane Davidson, "Improving Evaluation Questions and Answers: Getting Actionable Answers for Real-World Decision Makers" (demonstration session at the 2009 Conference of the American Evaluation Association, Orlando, FL, November 18, 2009), http://comm.eval.org/resources/viewdocument/?DocumentKey=e5bac388-f1e6-45ab-9e78-10e60cea0666.

- Innovative interventions and pilot programs that, if proven successful, can be scaled up or replicated.
- Interventions where there is not a good understanding of their impacts, and better evidence is needed to inform decisions about whether to continue funding them or to redirect funding to other interventions.
- Periodic evaluations of the impact of a portfolio of interventions in a sector or a region to guide policy, future intervention design and funding decisions.
- Interventions with a higher risk profile, such as a large investment (currently or in the future), high potential for significant negative impacts or sensitive policy issues.
- Interventions where there is a need for stakeholders to better understand each others' contributions and perspectives.

The timing of an impact evaluation is important. If it is done too soon, there may be insufficient evidence of impacts having occurred or being sustained. If it is done too late, it can be difficult to follow up with participants and too late to influence decisions about the future direction of the intervention. In any case, it is better to plan the impact evaluation where possible from the beginning of the intervention. This allows for evidence to be gathered throughout the intervention, including baseline data, and allows the option of using methods like randomized controlled trials, which require creation of a randomly allocated control group from the beginning of implementation.

13. What Is Needed for Quality Impact Evaluation?

It can be helpful to think about quality evaluation in terms of five competing imperatives: *utility*, *accuracy*, *ethics*, *practicality* and *accountability*.[9] These five standards are often in tension – for example, a more comprehensive impact evaluation that will be more accurate might not be practical in terms of available resources, might be too intrusive in the data collected, or might take too long to complete for it to inform key decisions about the future of the intervention.

Utility – good impact evaluation is useful. The likely utility of an evaluation can be enhanced by planning how it will be used from the beginning, including linking it to organizational decision-making processes and timing, being clear about why it is being done and who will use it, engaging key stakeholders in the process, and then choosing designs and methods to achieve this purpose.

Accuracy – good impact evaluation is rigorous. It pays attention to all important impacts, noticing if any are unintended. It pays attention to the distribution of impacts, noticing if only some people benefit, and who those people are. Accuracy requires the use of appropriate evidence, including quantitative and qualitative data, appropriate interpretation, and transparency about the data sources that have been used and their limitations. Strategies to achieve accuracy include systems for checking the quality of the data at the point of collection and during processing, and that findings have been reported fairly, comprehensively and clearly.

Propriety (ethics) – ethical issues need to be adequately addressed – including confidentiality and anonymity, as well as potential harmful effects of being involved in the evaluation. Some ethical issues,

[9] Joint Committee Standards for Educational Evaluation http://www.jcsee.org/program-evaluation-standards/program-evaluation-standards-statements. These were originally developed for educational evaluation but are often widely used more broadly.

such as the need to honor promises made about privacy and confidentiality, are common across different types of evaluations and research. There are other issues that are particular to impact evaluation. Concerns are sometimes raised about the ethics of using an RCT design, as it involves withholding an intervention from some people (the control group). There is less ethical concern when access to the intervention is going to be rationed in any case, and can be addressed by allocating the control group to a queue so they do receive the intervention after the evaluation of the first phase has finished (if it is shown to be effective). However, this strategy is only feasible when the impacts (or credible predictors of them) will be evident early, and when the intervention will still be relevant for the control group by the time the evaluation has ended.

There are also potential ethical issues in terms of whose interests are served by an evaluation. The American Evaluation Association discusses this in its Guiding Principles in terms of "Responsibilities for General and Public Welfare":

> *Evaluators articulate and take into account the diversity of general and public interests and values, and thus should:*
>
> 1. *Include relevant perspectives and interests of the full range of stakeholders.*
> 2. *Consider not only immediate operations and outcomes of the evaluation, but also the broad assumptions, implications and potential side effects.*
> 3. *Allow stakeholders access to, and actively disseminate, evaluative information, and present evaluation results in understandable forms that respect people and honor promises of confidentiality.*
> 4. *Maintain a balance between client and other stakeholder needs and interests.*
> 5. *Take into account the public interest and good, going beyond analysis of particular stakeholder interests to consider the welfare of society as a whole.*

Formal approval by the appropriate institutional review board is usually needed to undertake an impact evaluation. Applications for approval need to follow the prescribed format and address issues of beneficence, justice, and respect. (Evaluation Consent and the Institutional Review Board Process)

Practicality – impact evaluations need to be practical. They must take into account the resources that are available (time, money, expertise and existing data) and when evaluation results are needed to inform decisions. Partnering with one or more evaluation professionals, research organizations, universities and civil society organizations can leverage the necessary resources.

Accountability – evaluations need to make clear the evidence and criteria on which conclusions have been drawn, and acknowledge their limitations. Transparency about data sources is important, including showing which sources have been used for which evaluation questions. A formal process of meta-evaluation – having your own evaluation evaluated by approving an evaluation plan and then an evaluation report – by an expert reviewer or a committee of individuals with respected integrity and independence, can improve the accountability of an impact evaluation.

14. Common Challenges in Impact Evaluation in Development

A number of common challenges for development evaluation are described below, along with some suggestions for addressing them.

- **Variation in implementation and environment across different sites**
  An intervention may have been implemented in quite different ways to suit the different contexts in different country offices around the world, or in different geographic areas within a country. It can be useful to compare the theories of change for each site. In particular, identify whether different sites are using the same theory about how change happens (e.g., by increasing people's knowledge about their entitlements to services) but different action theories (e.g., printed brochures vs. community theater), or whether they are using different change theories altogether (e.g., increasing people's knowledge about their entitlements to services in one site vs. reducing barriers to service access – such as user fees – through advocacy in another).

- **Heterogeneous impacts**
  Development interventions often only work well for some people, and may be ineffective or even harmful for some other people. In addition, the success of an intervention in terms of achieving desirable impacts is often affected by the quality of implementation. It is therefore important to not only calculate and report on the average effect but to also check for differential effects. This requires gathering evidence where possible about the quality of implementation and data about contextual factors that might affect impacts, including participant characteristics and the implementation environment.

- **Diverse components**
  A program might encompass a diverse range of projects, and yet an overall evaluation of the impact of the whole program is needed. It can be helpful to develop an overall theory of change for the program, bringing together different components. Sometimes it is possible to do this in the planning stage, but, especially where projects or components have varied over time, this might need to be done retroactively.

- **Long time scales**
  Often the intended impacts will not be evident for many years, but evidence is needed to inform decisions before then (e.g., on whether or not to launch a subsequent phase or replicate the model elsewhere). A theory of change can identify intermediate outcomes that might be evident in the life of an evaluation. In some cases, research evidence can be used to fill in later links, and estimate likely impacts given the achievement of intermediate outcomes. Consideration should also be given to the expected trajectory of change – when impacts are likely to be evident. (Michael Woolcock on The Importance of Time and Trajectories in Understanding Project Effectiveness)

- **Influence of other programs and factors**
  The impacts of development interventions are heavily influenced by the activities of other programs and other contextual factors that might support or prevent impacts being achieved. For example, cash transfers that are conditional on school attendance will only lead to improved student achievement in situations where schools are teaching students adequately. It is possible to identify these other programs and contextual factors as

part of developing a theory of change, to gather evidence about them and to look for patterns in the data.

- **Resource constraints**
  Existing evidence (in the form of program documentation, baseline data and official statistics) may have gaps, and there may be few resources (in terms of funding, staff time or access to specialist technical expertise) to collect the types of evidence needed for quality impact evaluation. For a specific evaluation, when existing evidence is scarce and there are few resources to gather additional evidence, key informant interviews from diverse informants may provide sufficient data, including reconstructing baseline data. Planning ahead for impact evaluation can reduce resource constraints by building in sufficient resources at the design and budgeting stage, and/or strategically allocating evaluation resources across interventions so that they are concentrated more on a smaller number of more comprehensive evaluations of strategically important interventions.

Summary

An impact evaluation should begin with a plan that clarifies its intended purposes, including identifying intended users, the key evaluation questions it is intended to answer, and how it will address the six components of impact evaluation – clarifying values, developing a theory of change, measuring or describing important variables, explaining what has produced the impacts, synthesizing evidence, and reporting and supporting use. Having this plan reviewed (including by intended users) will increase the likelihood of producing a high quality impact evaluation that is actually used.

References and other Useful Resources

Alton-Lee, A. (2003) "Quality Teaching for Diverse Students in Schooling: Best Evidence Synthesis." Wellington, New Zealand: Ministry of Education. http://www.educationcounts.govt.nz/publications/series/2515/5959 *An example of synthesizing evidence from diverse sources to understand what works for whom.*

Catley A., Burns, J., Abebe, D., Sufi, O. Participatory Impact Assessment: A Guide for Practitioners. Boston: Tufts University. http://www.preventionweb.net/english/professional/publications/v.php?id=9679

Chambers R. (2007) "Who Counts? The Quiet Revolution of Participation and Numbers" Working Paper No. 296, Brighton: Institute of Development Studies. http://www.ids.ac.uk/files/Wp296.pdf

Davidson, E. J. (2006) "Causal Inference Nuts and Bolts." Demonstration session at the 2006 Conference of the American Evaluation

Association, Portland, Ore., Nov. 2006 http://realevaluation.com/pres/causation-anzea09.pdf.

Davidson, E. J. (2009) "Improving Evaluation Questions and Answers: Getting Actionable Answers for Real-World Decision Makers." Demonstration session at the 2009 Conference of the American Evaluation Association, Orlando, Fla., Nov. 2009. http://comm.eval.org/resources/viewdocument/?DocumentKey=e5bac388-f1e6-45ab-9e78-10e60cea0666.

Funnell S. and Rogers, P. (2011) *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. San Francisco: Jossey-Bass/Wiley.

Guijt, I. (1999) Participatory Monitoring and Evaluation for Natural Resource Management and Research. Socio-economic Methodologies for Natural Resources Research. Chatham, UK: Natural Resources Institute. http://www.nri.org/publications/bpg/bpg04.pdf

Miles, M. and Huberman, M. (1994) *Qualitative Data Analysis: An Expanded Sourcebook* (2nd ed.) Thousand Oaks California; Sage Publications. *Outlines strategies for checking causal explanations, including searching for disconfirming evidence, following up exceptions, making and testing predictions.*

Patton, MQ (2008) "State of the Art in Measuring Development Assistance." *Discuses the importance of interpretation and managing uncertainty in effective management.*

Paz R., Dorward A., Douthwaite B. (2006). "Methodological Guide for Evaluation of Pro-Poor Impact of Small-Scale Agricultural Projects." Centre for Development and Poverty Reduction. Imperial College, London. http://boru.pbworks.com/f/modulosjan07.pdf *Describes 22 methods and tools that can be used to evaluate the direct and indirect impacts of innovation adoption.*

Roche, C. (1999) Impact Assessment for Development Agencies: Learning to Value Change Oxford: OXFAM, Novib

Rogers, Patricia J. (2009) "Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation", *Journal of Development Effectiveness*, 1(3): 217-226. Working paper version available at: http://www.3ieimpact.org/admin/pdfs_papers/50.pdf.

Walter, S., Olivier, J., Churches, T. and Grzebieta, R. (2011). "The impact of compulsory cycle helmet legislation on cyclist helmet head injuries in New South Wales, Australia", *Accident Analysis and Prevention*, 43 : 2064-2071.

White, S. and J. Petit (2004) Participatory Methods and the Measurement of Wellbeing Participatory Learning and Action 50, London: IIED

www.betterevaluation.org – information on evaluation methods for development, including user-contributed examples and comments

www.mymande.org – information, videos and links to information about evaluation methods