Field Epidemiology
In Action

Impact Evaluation

James Flint

Introduction to Impact Evaluation

While ‘impact’ is typically included in models used for evaluating training programs, often under the terms ‘results’ and ‘return on investment’, the evolution of impact evaluation methods has largely come from the development sector. International development agencies are placing increasing emphasis on addressing development effectiveness and impact. (1) For many years, evaluation experts have focused on measuring program outputs and outcomes. (2) Given the high monetary and opportunity costs of many training programs, the move towards demonstrating value for money and measuring impact is a profoundly reasonable expectation. Demands for increased accountability and transparency come not only from donors, but also from those impacted by programs – often the poor and marginalized in development-related projects – who want to know how they and their communities will benefit following their engagement with development partners. (3)

In 2012, DFAT’s Impact Evaluation Working Group released a discussion paper for AusAID practitioners. (4) This paper outlines the following four reasons for undertaking impact evaluations:

1. To establish the value of innovative and effective programs.
2. To test the causal logic and assumptions of an intervention that has long been used, but has little and/or contested evidence on how it contributed to development outcomes.
3. To test the validity of a successful intervention in a different context.
4. To prove the worth of an intervention to policy-makers and decision-makers.

DFAT defines impact evaluation as “a systematic and empirical investigation of the impacts produced by an intervention – specifically, it seeks to establish whether an intervention has made a difference in the lives of people”. (4) Impact evaluation seeks to demonstrate that intended results follow from program activities directly or indirectly. (3) It addresses questions as to what works and does not work, how, why, and for whom. In order to do this, impact evaluations link cause and effect, assessing the direct and indirect causal contribution of the intervention. It includes both positive and negative, intended and unintended, direct and indirect, primary and secondary effects resulting from an intervention. (5, 6)

Rogers states that when choosing impact evaluation methods, the following six aspects of an impact evaluation should be addressed: (6)

1. Clarifying the values that will underpin the evaluation – what will be considered desirable and undesirable processes, impacts, and distribution of costs and benefits?
2. 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).
3. Measuring or describing these impacts and other relevant variables, including processes and context.
4. Explaining whether the intervention was the cause of observed impacts (casual attribution or plausible contribution).
5. Synthesizing evidence into an overall evaluative judgment.
6. Reporting findings and supporting their use.

Determining and testing causal attribution is a requirement for impact evaluations. The design options for testing causal attribution include experimental, quasi-experimental, and non-experimental.

Common strategies and methods to establish the causal attribution or plausible contribution are: (4, 6)

- ***Factual assessment*** (best addressed by theory-based models)– the extent to which the actual results match what was expected; is what was observed in the program/intervention and the broader environment consistent with the theory:
    - Comparative case studies: did the program/intervention produce results only when the necessary elements were in place.
    - Dose-response: were there better outcomes for participants who received more of the program/intervention.
    - Beneficiary/expert attribution: did participants/stakeholder believe the program/intervention made a difference.
    - Predictions: did participants or sites predicted to achieve best outcomes/impacts do so.
    - Temporality: did the outcomes/impacts occur at a time consistent with the theory of change (note, theory of change is described below).

- ***Counterfactual assessment*** (best addressed by experimental and quasi-experimental models) – an estimate of what would have happened without the program or intervention:
    - Difference-in-difference: before and after difference for the group receiving the program/intervention is comparted to the before and after of those who did not.
    - Logically constructed counterfactual – a before and after comparison with the group receiving the program/intervention.
    - Matched comparison (case-control): participants (individuals, organizations, communities) are matched with nonparticipants on variable thought to be relevant.
    - Multiple baselines: staggered implementation of program/intervention over time; analysis of repeated patterns.

    - Propensity scores: statistical creation of a comparable group based on factors that influenced people’s propensity to participate in a program/intervention.
    - Randomized controlled trial: participants (individuals, households, communities) are randomly assigned to receive the intervention or be in a control group.

- ***Rule out possible alternative explanations*** – (best addressed by theory-based models) identifying and testing plausible alternative explanation for the changes:
    - General elimination methodology: possible alternative explanations are identified then investigated.

Impact evaluation has moved away from sole dependence on experimental designs. Experimental designs are concerned with intended rather than unintended effects; assume direct links between interventions and outcomes; address primary rather than secondary effects; and usually look to short term evidence rather than long term. (5) Understanding why and how programs succeed or fail, in order to improve or replicate them with confidence, has become an increasingly important reason for conducting impact evaluations and has led to the rapid development and adoption of non-experimental evaluations.

Impact Evaluation Methodologies

Impact evaluation methodologies generally fall into experimental/quasi-experimental and non-experimental categories. Experimental and quasi-experimental designs provide quantitative measures of the net effect of a program. Theory-based models are a common non-experimental approach. They provide insights required to improve, replicate, and scale activities and are best suited for evaluations with small sample sizes. (1)

Experimental and quasi-experimental

Built on a reductionist theoretical foundation, experimental and quasi-experimental designs assume linear causal relationships between program elements and desired program outcomes. While experimental designs have been enormously useful in advancing the biological sciences, they have proven less useful in the evaluation field. (7) The tightly controlled designs of experimental and quasi-experimental methods are typically very difficult to implement for complex training and education-based impact evaluations. (8) Neither experimental nor quasi-experimental designs are good at dealing with contextualization – taking account of cultural, institutional, historical, and economic settings.

Experimental and quasi-experimental methods compare performance in a group receiving the intervention with performance in a group not receiving the intervention. The impact is measured by comparing the difference in the intervention and control groups. Experimental and quasi-experimental methods require the identification of suitable control groups and usually require large surveys. They are good as answering the question: ‘has this particular intervention made a difference here?’, but are weak at answering generalisation (external validity) questions: ‘will it work elsewhere’. Experimental designs randomly allocate the intervention (‘treatment’) in the population and test how well it achieves its objectives, as measured by a pre-specified set of indicators. Impact evaluations based on experimental methods are referred to as randomized controlled trials. Quasi-experimental methods, by definition, lack random assignment. The assignment to the intervention and non-intervention groups is determined by the participants (self-selection) or the evaluators. (9) Quasi-experimental methods identify a control group as similar as possible to the treatment group in terms of pre-intervention characteristics. There are different methods for creating a valid control group, including propensity score matching and regression discontinuity design. Quasi-experimental designs usually use existing data and are usually cheaper to implement than experimental designs.

Theory-based (non-experimental)

It is often helpful to base an evaluation on a theory or model of how the program or intervention is understood to produce the intended impacts. This is especially important when assessing the contribution of program or intervention on the intended outcome and anticipated impacts. These theories or models are often referred to as program theory, theory of change, results chain, or logic model. (6) Theory-based approaches are a “logic of enquiry” which complement and can be used in conjunction with many evaluation designs. (10) Theory-based approaches attempt to understand a program's contribution to observed results through a mechanistic or process interpretation of causation rather than determining causation through comparison to a counterfactual.

There are several different approaches to theory-based impact evaluations. The common factor is that they all rely on identifying the mechanism that explains effects. They also share two key stages. First, a conceptual stage where researchers work with local stakeholders to develop the causal mechanism used to guide the evaluation. Second, an empirical stage where researchers test the causal mechanisms to bring about the observed outcomes. (11)

Theory-based evaluations not only describe the outcomes and impact of a given program but also provide an understanding of the program’s role in producing them. It is rarely the case that a program or intervention is the sole cause of the changes observed in the evaluation. There are often many other factors at play, influencing the results both positively and negatively. A theory-based approach is validated through empirical evidence to test its underlying assumptions and hypothesis that represent alternative causal explanations. Validating a theory reduces uncertainty about the contribution a program made to the observed outcomes and impact. The level of confidence in the causal link depends on the method used for testing this link. Having a theory permits evaluators to determine where unsuccessful programs broke down and, thus, provide valuable intelligence on where to focus improvements. In the event of a program failure, a theory will also allow evaluators to distinguish between implementation failure (program was not implemented properly) and theory failure (program was implemented properly, but still failed). A theory-based approach can also help manage potential negative impacts. (6) By highlighting critical steps that require improvement or those that are unnecessary in bringing about change, theory-based evaluations promote cost effective practice and are useful in understanding social and power structures that may influence causal links. (11)

The two key theory-based impact evaluation approaches are the Theory of Change and Realistic Evaluation. Developing the **theory of change** is done with the active participation of the local stakeholders to ensure the process is open to different perspectives and insights. (11) Hivos defines theory of change as “the ideas and hypotheses people and organisations have on how change happens.” (12) Theories of change can improve evaluation by helping identify intermediate outcomes or impacts that can be observed within the timeframe of the evaluation. These 'impact indicators' are precursors to the longer-term impact that the program was designed to achieve. This is especially relevant to health-related training initiatives where long-term impacts may occur over many years. Once the theory of chance has been developed, evaluation is done using various tools to measure and understand its impact. Many of the tools used are case-based approaches and utilize qualitive methods; examples include:

- **Case-based evaluation** (case-studies and stories of change): this approach focuses on the systematic generation and analysis of case studies or stories of change. They are often used as alternatives or supplementary to quantitative reporting. The within-case analysis considers an individual case in detail and describes what changed, how the change came about, the contribution of the intervention, the contribution of other factors, and the lessons learned. The cross-case analysis assesses change drawn from multiple cases. Case-based analysis is often appropriate when evaluating complex interventions. (3) They can cope with complex change as they do not rely on pre-defined indicators. However, case-based evaluations are not always the best approach if there is a need is to generate universal findings that can be applied in different situations or locations. Stories of change are similar to case studies; however, they are always focused on chance. Stories of chance usually attempt to show how an intervention has continued to the desired change.

  There are a number of different ways to categorise case studies used for evaluation; these include:
  - **Best cases** – used to showcase the biggest and most important changes.
  - **Typical cases** – used to describe the typical effect of an intervention.
  - **Illustrative cases** – used to illustrate a key point of the message.
  - **Comparative cases** – used to compare between two or more different situations or to compare chance across different individuals or settings.
  - **Learning cases** – used to communicate significant learning that can be used to improve performance within an intervention or more widely.

  Although case studies or stories of change may be used on their own, it is also common to include a set of cases chosen according to specific criteria. In qualitative analysis, this is referred to as purposeful sampling. Examples of purposeful sampling criteria include unusual, extreme or deviant cases, homogenous cases, and criterion sampling (used to investigate cases that meet specific criteria). (13, 14)

- **Qualitative Comparative analysis**: this method incorporates a methodology which enables the analysis of multiple cases on complex situations. It can help explain why change happens in some cases but not others. The first step is to develop a theory of change, followed by identifying cases of interest, developing a set of evaluation factors, scoring those factors, and analyzing and interpreting the results in light of the theory of change.

- **Most significant change**: this technique is a form of participatory evaluation. It involves the collection and selection of stories of change produced by program stakeholders. It was first developed in Bangladesh in the 1990's, and since has been used by a variety of organizations. The process begins with defining domains of change, deciding how and when to collect stories, collecting significant change stories, selecting the most significant stories and verifying the stories. The most significant change is not designed to provide a comprehensive overview of the change resulting from an intervention or program; it is not designed to describe typical change, rather, the most significant change. If typical change is needed, this method needs to be complemented by other methodologies. (15, 16)

- **Success case method**: the success case method, is an evaluation process developed by Robert Brinkerhoff in 2003, and it involves identifying the most and least successful cases in a program and examining them in detail. (17, 18) It is a useful approach to document stories of impact and to develop an understanding of the factors that enhance or impede impact. The methods deliberately focus on the most, and least, successful participants of a program. Although it has been used in multiple settings, this method was originally developed to evaluate the impact of training interventions. The approach relatively timely and straightforward, providing an attractive alternative to the more complex experimental and realist approaches. It is considered an ideal evaluation tool to complement training evaluation frameworks such as Kirkpatrick. The success case method allows inquiry beyond the narrow 'training alone' focus of Kirkpatrick - and related approaches - allowing consideration of the broader performance context. The success case method does not seek to isolate the effect of training, but instead, draws on performance-systems thinking that acknowledges the inseparability of learning and performance. (18) The success case method begins by identifying likely success cases – individuals or teams who have been the most successful in applying learning from a training program, for example. Identification of success cases often occurs through a survey. The second part of the method involves interviewing cases and documenting evidence of success. The success case method combines the craft of storytelling with more current evaluation approaches of naturalistic inquiry and case study. It also employs the social inquiry process of key informants and borrows tools from journalism and judicial inquiry.

The **realist evaluation** approach was first developed by Pawson and Tilley in 1997 and has been adapted in many different ways since. (19) Realist evaluation is rooted in the philosophy of realism and are based on the assumption that projects and programs work under certain conditions, and are heavily influenced by the way that different stakeholders respond to them. Understanding how and why projects and programs work in a different context is an important focus for realist evaluation. Realist evaluation is designed to address questions such as 'what works, for whom, in which circumstances and how and why does it work'. It is heavily focused on causation – assessing which initiatives contributed to different results and how. As such, realist evaluations are particularly useful in understanding why, how, and for whom a project or program works. They are useful for evaluating interventions that are intended to be expanded, replicated, or scaled up, or for understanding inconsistent results from prior evaluations. (20) Realist evaluation theories are usually based on the context-mechanism-outcome (CMO) hypothesis. This, in practice, means theorizing different outcomes to interventions in different contexts. This is opposed to normal program theory, which tends to assume that changes at one level lead to further changes at higher levels irrespective of the context. Realist evaluation is often based on multiple case studies or stories of change with outcome data being disaggregated to allow comparisons between different groups and sub-groups. Realist evaluation is complex, and the skills required to undertake realist evaluation are more significant than many other kinds of evaluations. Therefore, a realist evaluation should only be undertaken when there is a convincing case for carrying it out that justifies a larger investment. It is not something that is designed to be applied lightly or cheaply. (20)

DFAT, in line with good international practice, recommends the following minimum standards for impact evaluations: (4)

- ***Mixed approaches and methods***, incorporating both a mix of evaluation approaches and integrated analysis of both quantitative and qualitative data.
- ***Theory-based***, to identify and test causal pathways.
- ***Analysis of sub-populations***, including different impacts on poor people, men and women, boys and girls.
- ***Systematic collection and analysis of quality data***.
- ***Appropriately resourced***.

The impact evaluation design chosen must be grounded in the key evaluation question being asked. There are four main questions impact evaluators ask; they are listed in table 3 with the corresponding underlying assumptions and suitable designs. (3) As evaluators often choose to answer more than one of these key evaluation questions, mixed designs, as well as mixed methods, are often required.

Table 3. Design implications of different impact evaluation questions.

Key evaluation questions	Related evaluation questions	Underlying assumptions	Requirements	Suitable designs
To what extent can a specific (net) impact be attributed to the intervention?	What is the net effect of the intervention? How much of the impact can be attributed to the intervention? What would have happened without the intervention?	Expected outcomes and the intervention itself clearly understood and specifiable Likelihood of primary cause and primary effect Interest in particular intervention rather than generalisation	Can manipulate interventions Sufficient numbers (beneficiaries, households etc) for statistical analysis	Experiments Statistical studies Hybrids with case-based and participatory designs
Has the intervention made a difference?	What causes are necessary or sufficient for the effect? Was the intervention needed to produce the effect? Would these impacts have happened anyhow?	Several relevant causes need to be disentangled Interventions are just one part of a causal package	Comparable cases where a standard set of causes are present and evidence exists as to their potency	Experiments Theory-based evaluation Case-based designs Contribution Analysis Success Case Method process

How has the intervention made a difference?	How and why have the impacts come about? What causal factors have resulted in the observed impacts? Has the intervention resulted in any unintended impacts? For whom has the intervention made a difference?	Interventions interact with other causal factors It is possible to clearly represent the causal process through which the intervention made a difference – may require 'theory development'	Understanding how supporting and contextual factors that connect intervention with effects Theory that allows for the identification of supporting factors – proximate, contextual and historical	Theory-based evaluation especially 'realist' variants Contribution Analysis Success Case Method process Participatory approaches
Can this be expected to work elsewhere?	Can this 'pilot' be transferred elsewhere and scaled up? Is the intervention sustainable? What generalisable lessons have we learned about impact?	What has worked in one place can work somewhere else Stakeholders will cooperate in joint donor/ beneficiary evaluations	Generic understanding of contexts e.g. typologies of context Clusters of causal packages Innovation diffusion mechanisms	Participatory approaches and some Experimental and Theory-based approaches Realist evaluation

Adapted from Stern, 2015. (3)

Conclusion

The development and expansion of FETPs over the past 70 years demonstrates the value and appeal of this applied on-the-job training model. The compilation of outputs achieved by fellows is long and impressive. There remains, however, a gap in the literature documenting the *impact* of FETPs. As the public health landscape continues to shift, and especially in light of COVID-19, the need to examine FETPs in order to appreciate whether, how and why they are making a difference is critical. The case for developing a national workforce capable of preparing for, detecting and responding to emerging disease threats hardly needs mention in light of COVID-19. The case on how best to do this is less clear.

As both a training and a development program, the evaluation of FETPs will benefit from a mixed model and mixed methods approach. There are several tried and tested evaluation models that can be used together to extend the reach of FETP evaluations beyond processes and outputs, to a careful examination of program outcomes and impacts. As new FETP programs are implemented and existing training models developed, delving into the world of impact evaluation is long overdue.

References

1. White H and Phillips D. Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework. 3ie Working Paper 15 [Internet]. 2012 27 Jan 2020. Available from: https://www.3ieimpact.org/evidence-hub/publications/working-papers/addressing-attribution-cause-and-effect-small-n-impact.

2. Patton M. Developmental evaluation: applying complexityconcepts to enhance innovation and use. New York: Guilford Press; 2011.

3. Stern E. Impact Evaluation. A giude for commisioners and managers. In: Development DfI, editor. United Kingdom2015.

4. Australia Department of Foreign Affairs and Trade. Impact Evaluation: A Discussion Paper for AusAID Practitioners2012 27 Jan 2020. Available from: https://dfat.gov.au/aid/how-we-measure-performance/ode/Documents/impact-evaluation-discussion-paper.pdf.

5. Niels D. Glossary of key temrs in evaluation and results based management. France: OECD Working Party on Aid Evaluation; 2010.

6. Rogers P. Introduction to Impact Evaluation. Impact Evaluation Notes No. 1. Washington: InterAction; 2012.

7. Stufflebeam DL. CIPP Evaluation Model Checklist, A tool for applying the CIPP Model to assess long-term enterprises: Evaluation Checklists Project 2007 [cited 2020 Jan 7]. Available from: https://wmich.edu/sites/default/files/attachments/u350/2014/cippchecklist_mar07.pdf.

8. Frye AW, Hemmer PA. Program evaluation models and related theories: AMEE guide no. 67. Medical Teacher. 2012;34(5):e288-99.

9. Shadish WR, Cook TD, Campbell DT, editors. Experimental and Quasi-Experimental Designs for Generalized Causal Inference2001.

10. Treasury Board of Canada Secretariat. Theory-based approaches to evaluation: concepts and pracxtices. Ottawa, Canada: Centre for Exxcellence for Evaluation; 2012.

11. Mohammed EY and Bladon A. Theory-based impact evaluation. International Institute for Environment and Development; 2017 March.

12. Marjan van Es IG, Isabel Vogel. Theory of Change Thinking in Practice: A stepwise approach. In: Hivos, editor. 2015.

13. Roche C. Impact Assessment for Development Agencies. 1999.

14. M P. Qualitative Evaluation and Research Methods. London; 1990 1990.

15. Davies R. An evolutionary approach to facilitating organisational learning: an experiment by the Christian Commission for Development in Bangladesh. Impact Assessment and Project Appraisal. 1998;16(3):243-50.

16. Davies R. The ‘Most Significant Change’ (MSC) Technique: A Guide to Its Use”2015.

17. R.O. B. The success case method: find out quickly what’s working and what’s not. San Francisco: Berrett-Koehler; 2003.

18. Brinkerhoff R. Training The Success Case Method: A Strategic Evaluation Approach to Increasing the Value and Effect of. Advances in Developing Human Resources. 2005;7.

19. Pawson RaT, N. Realistic Evaluation. London: SAGE; 1997.

20. Westhorp G. Realist Impact Evaluation. An Introduction. . ODI; 2014 September 2014.