Journal of MultiDisciplinary Evaluation
Volume 12, Issue 27, 2016

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Debate on the Appropriate Methods for Conducting Impact Evaluation of Programs within the Development Context

Enyonam B. Norgbey
University of Ottawa

Background: Donors and decision-makers use impact evaluation reports to assess the effectiveness of development programs and identify ways to improve the design and implementation of projects, programs, and policies in developing countries.

Purpose: This paper will explore the existing published impact evaluation literature on development programs and provide an overview of the types of approaches and methods that are being used to conduct impact evaluations.

Setting: NA

Intervention: NA

Research Design: The paper will examine published program evaluation literature in order to shed light on issues related to appropriate methods for impact evaluations of development programs.

Data Collection and Analysis: Literature review.

Findings: The paper will conclude by suggesting a list of approaches and methods that can be used to conduct impact evaluations of programs within the development context.

Keywords: *impact evaluation; evaluation methods; development programs; international development; development effectiveness.*

Introduction

For the past twenty years, the international development community has emphasized the use of impact evaluation to assess the effectiveness of projects, programs, and policies (Cameron, Mishra, & Brown, 2016). The global trend in impact evaluation has been steered by various economic and political factors, and many donors have advocated for accountability and more evidence of 'value for money' (White, 2010). Moreover, a fierce campaign for evidence-based policy during the last decade has gained strength leading to the scrutiny of some of the presumptions made regarding the efficacy of development assistance programs [1] (Hearn & Buffardi, 2016). To that effect, programs in Third World countries in general, and Africa, in particular, have focused on a broader vision of development[2]. This approach is different from the 'project based' approach of the mid-1990s, where development assistance was provided in the form of donor 'owned' projects, with deeply constrained sets of activities relying on deliverables supported by a 'logical framework' (Conlin & Stirrat, 2008). Moreover, the current complex environment of development programs has been exacerbated by the (MDGs)[3], because donors and stakeholders have generally emphasized holistic approaches to evaluating program results (Conlin & Stirrat, 2008), with focus on collaborative and partnership processes. According to (Lilja et al., 2010), this holistic approach helps build institutional and technical capacity while making use of local knowledge and capacity. Meanwhile, perspectives on the definition, scope, and appropriate methods for conducting impact evaluations vary among practitioners and stakeholders. While current debates on impact evaluation have involved various issues, the debates have been most intense on the methods of research, research design, and data collection (NONIE, 2009).

The subject of impact evaluation is of interest to evaluation practitioners because the use of rigorous evaluation methods will: on the one hand, provide the full picture on the successes or failures of development programs, and on the other hand, help improve the design of projects, programs, and policies. More importantly, it is my view that the evaluation of development outcomes must produce tangible and measurable evidence, and the most appropriate tool for program assessment is impact evaluation. In this paper, existing published program evaluation literature will be explored and an examination of the types of approaches and methods used in conducting impact evaluations will help shed light on issues related to impact evaluation methods. The definition of impact evaluation will first be discussed, followed by various types of evaluation approaches and methods used for impact evaluations. Finally, the conclusion will present the author's reflections and position on the impact evaluation debate related to development programs in Third World countries in general, and in Africa in particular.

What is Impact Evaluation?

Impact evaluation is one of the ways to determine whether development programs brought about the expected change; whether the program or project should be discontinued, scaled down or expanded; whether there were unexpected positive or negative results; and if so, to determine what the causes were (Hearn & Buffardi, 2016). Above all, impact evaluation imparts the information necessary to provide accountability to donors, stakeholders, and beneficiaries as well as the lessons learned that will inform future programs (Hearn & Buffardi, 2016).

As previously outlined, the definition of impact evaluation varies widely among donors and various stakeholders in the development community. There are two major ideologies about the meaning of impact evaluation. According to the Development Assistance Committee (DAC) of the Organization for Economic Cooperation and Development (OECD), impact evaluation is "the positive and negative, primary and secondary long-term effects produced by a development program, directly or indirectly, intended or

[1] Development assistance is a financial aid provided by governments and other agencies to promote the economic, environmental, social and political development of developing nations. Retrieved from Wikipedia on April 7, 2016. https://en.wikipedia.org/wiki/Development_aid.

[2] "Development is the process of economic and social transformation that is based on complex cultural and environmental factors and their interactions." Retrieved from the online Business Dictionary on April 7, 2016, from http://www.businessdictionary.com/definition/development.html.

[3] "The eight Millennium Development Goals (MDGs) – which range from halving extreme poverty rates to halting the spread of HIV/AIDS and providing universal primary education, all by the target date of 2015 – form a blueprint agreed to by all the world's countries and all the world's leading development institutions.." (UN website. Retrieved on April 7, 2016: http://www.un.org/millenniumgoals/)

unintended" (OECD/DAC, 2002). The World Bank, on the other hand, defines impact evaluation as the appraisal of the transformation that occurred in the welfare of the people, the household, the communities, and the regions as a result of a particular program. The emphasis is on the causality or the "attribution" of the transformation and what would have happened in the absence of the program, the "counterfactual" (World Bank, 2011). Given the fact that each of the two definitions involves specific outcomes, a variety of methods will be favored by the respective institutions. Hearn and Buffardi (2014) assert: "the lack of consistent definition and technical debates about methods have led to confusion among donors and implementation staff". This is one of the causes of the current controversy over methods selection to conduct impact evaluation within the development context.

The Debate

Since the international development community cannot agree on a single definition of impact evaluation, it is not surprising that some groups advocate for quantitative methods while others call for mixed-methods approaches. Carvalho and White (2004) argue that quantitative approaches and the use of randomized controlled trials (RCTs) with experimental and quasi-experimental designs are the best available methods for conducting a rigorous impact evaluation with bias-free participant selection. White (2010) posits that attribution and counterfactual analyses provide information related to what works and also allows cost-benefit analysis. He further claims that RCTs are a 'gold standard'. Others assert that their approach to impact evaluation of development programs is 'situational responsiveness' (Rogers, 2009; Patton, 2008a)

Stame (2010) argues that the issue of the gold standard is the 'mother of all debates' for, it leads to more debates. She further adds that the development of a counterfactual effect is not necessarily the only approach that confirms causality. In fact, Patton (2008b) concurs that "the only methodological 'gold standard' is appropriateness of methods selected". The use of RTCs is limited in the real world and is only suitable in a few cases. For example, the impact of drugs on health and the impact of literacy on increased educational budget can be assessed by the RCT approach (Conlin & Stirrat, 2008). Conlin and Stirrat (2008) further argue that the increasing ambitious development goals, multiple layers of governance, and the variety of issues related to development trajectories are simply too complex to be assessed by basic cause and effect approaches. If one single approach is the only plausible way of getting impact evaluation, this implies that all programs must follow the same standard, which does not seem to be the case in the international development context. A review of published program evaluations will provide insight into specific approaches that are being used. This article will focus on three main approaches namely: quantitative (RCTs), theory-based, and mixed-methods approaches that require participatory methods.

Randomized Controlled Trial Approach

Even though RTCs[4] evolved two centuries ago, they have gained popularity recently due the Center for Global Development's (CGD) study in 2006 that advocated for the use of more RCTs in the evaluation of development programs with the expectation that they would provide a better understanding of 'what works well'. RCTs are a research approach that relies on two or more randomized groups where one is the "treatment" and the other is the "control" group. This design allows for the comparison of the control and the treatment groups at the end of the study to assess the effectiveness of a program on the treatment group (White & Sabarwal, 2002). This technique has been known to be very effective in analyzing the changes attributed to international development programs and providing accurate results if the control group has not been inadvertently exposed to the program as the treatment group (White, 2010).

A concrete example of RCT is described in a study undertaken by Beltramo and Levine (2013) to explore the effect of solar ovens on fuel use, time spent collecting wood, carbon monoxide exposure, and respiratory illness symptoms in rural Senegal. The objective of the intervention was to reach 2.5 per cent reduction in kilograms of wood used per cooking session, 3.5 per cent reduction in hours spent weekly to collect wood, and 9.8 per cent reduction in a cook's personal exposure to carbon monoxide while cooking. A randomized controlled trial was conducted with a sample made up of women interested in buying

[4] A randomized controlled trial is a type of scientific experiment that was "the first reported clinical trial conducted by James Lind in 1747 to identify treatment to vitamin c deficiency" Retrieved on May 7, 2016 from: https://en.wikipedia.org/wiki/Randomized_controlled_trial

solar ovens. Four hundred and sixty-five households randomly selected as part of the treatment group and 325 households as the control group completed a baseline survey. Households assigned to the control group were given their stoves six months after receiving training on how to use the solar stove. Eighty percent of the participants who generally cook far more than the volume of the solar oven continued using their traditional stoves and the solar oven during the time of the intervention. After six months of possession of the solar oven, a follow-up survey of the treatment group revealed that there was no significant reduction in fuel usage, cooking time, or time spent collecting wood.

The findings revealed that there was also no evidence that the solar ovens decreased exposure to carbon monoxide or self-reported respiratory symptoms like coughs and sore throats. Meanwhile, the evaluators argued that the evaluation was a policy success, since the goal was to determine whether the program could be expanded nationwide. Nevertheless, Beltramo and Levine acknowledged several shortcomings in the project implementation as well as the evaluation process. The RCT confirmed that the solar oven was a poor choice for the selected population. Moreover, the evaluators admitted that data collection was organized around 'household', ignoring the fact that in many houses, women take turns on cooking duties. In addition, there were various project design issues that required further consideration. For example: (1) there was a mismatch between the feasibility study and results of the RCT that could have been attributed to misaligned incentives for the stove design team; (2) the stove use monitors (SUMs) were not placed on the old stoves that were also being used during the intervention; (3) the carbon monoxide tubes designed to read the gas emission were placed only on the stoves of the women who intended to use their old stove (light fire) on specific days, while the ones cooking with gas were omitted; and (4) the carbon monoxide tubes measured just the carbon monoxide emission at mid-day without accounting for the emission during evening cooking. Finally, the evaluators relied solely on the women's self-reported use of wood and the time spent collecting wood that might not necessarily reflect the exact figures. Overall, the results of the RCT demonstrate that the solar oven did not meet the needs of the population given the fact that the women have complex cooking patterns requiring multiple stoves with multiple cooks.

The second RCT example is a study by Wang, Connor, Guo, Namboa, Chanda-Kapata, and Lambo et al. (2016) to evaluate the impact of non-monetary incentives to encourage health facility delivery in rural Zambia. The study was based on a clustered randomized controlled trial to measure the impact and cost-effectiveness of a four dollar 'Mama Kit' incentive[5] provided to mothers who delivered their babies in rural health facilities. The main objective was to reduce pregnancy and childbirth-related mortality in Zambia. The sample size was thirty facilities with an average of one hundred women per facility clustered into treatment and control groups. The evaluation was conducted to determine the effect of Mama Kits on facility delivery rates in the thirty rural health facilities selected. "The facility-level antenatal care and delivery registers were then used to measure the percentage of women attending antenatal care who delivered at a study facility during the intervention period" (p. 515).

The findings from the trial were then utilized to calculate the incremental cost-effectiveness ratio per death prevented due to the distribution of the incentives for all rural facility deliveries. The results revealed a 9.9 percentage point increase in the rural health facility delivery rate. The evaluation confirmed that the low-cost incentive packages were very useful and cost-effective ways to help increase rural facility delivery. However, the kits are unlikely to contribute to a comprehensive solution to safe delivery challenges. Moreover, the factors that influence whether a woman delivers at a facility are generally based on a complex combination of individual and situational matters. Even though the RCT approach was very relevant in quantifying the impact, it has not been able to answer all the questions related to 'why and how' the increase occurred.

RCTs have many limitations in their application to comprehensive programs with extensive scope at country, regional, or global levels as well as in various activities that cut across sectors, themes and geographic areas (Vaessen, 2010). As demonstrated in the above examples, Bamberger, Tarsilla, and Hesse-Biber (2016, p.156) concur that "many widely used evaluation designs, such as RCTs measure only intended consequences of an intervention and have significant methodological limitations regarding ability to identify unanticipated consequences" of an intervention. Consequently, many development agencies such as the World Bank and others call for more 'rigorous and relevant' impact evaluations with more contextualized and policy-

[5] 'Mama kit' is a small incentive package of child care items given mothers if they deliver their babies in rural health facilities in Zambia (Wang, et al., 2016).

relevant studies. Many researchers like Weiss, Carvalho, White, Sridharan, Nakaima, and others have used theory-based evaluation approaches to strengthen internal and external validity of results by explaining how and why certain changes happen (Vaessen, 2010). There was a growing demand for 'what works' and 'why'. To that effect, it is generally accepted that a theory-based approach to impact evaluation together with the examination of causal links from inputs to outcomes to impacts and the scrutiny of the undisclosed assumptions may be the best alternative to RCTs (White, 2009).

Theory-Based Evaluation Approach

The main goal of the theory-based approach to impact evaluation is to establish plausible causal links between the program and its impact (Coryn, Noakes, Westine, & Schröter, 2011). A number of theory-based approaches depend on a theory of change and a logical framework (Mayne, 2012). The overall goal is to minimize uncertainty of the contribution of the program to the observed change and the understanding of why the observed impacts occurred (NONIE, 2009).

The study of a maternal health service delivery in Kabale District in rural Uganda is a good example of the application of the Theory Based Evaluation (TBE) approach to impact evaluation even though it was combined with a Process Tracing Method[6] (PTM). Bamanyaki and Holvoet (2016) used a combination of the two approaches to explore the effects of local-level civil society-led gender-responsive budgeting. The study included four steps: (1) articulation of the program theory and how gender budget initiative would lead to the intended results; (2) hypothesizing the underlining causal process to describe the causal contribution to the observed and intended outcomes; (3) making predictions of observable exhibitions of the process; and (4) testing the empirical effects. The evaluation focused on the causal attribution based on the combination of program theory, experimental and non-experimental designs, to determine and assess the intermediary steps of program implementation, and the contribution of the intervention to the program.

The evaluation concluded that the use of the TBE-PTM model with an in-depth case study provided a plausible framework for the evaluation of gender mainstreaming programs that should lead to political, social, and cultural change. The evaluation further added that there was a political will among the stakeholders towards gender equality. However, the authors admitted that despite the political will, Kabale District is still waiting for an endorsed gender policy. Bamanyaki and Holvoet (2016) further stipulated that the lack of a district gender policy had not impacted on the operation of both mechanisms but was likely to negatively affect the sustainability of a gender perspective in health policies and budgets over time. Another alternative may be for the program to move towards affirmative actions to ensure gender equality and equity.

A second example of a theory-based impact evaluation is the Hombrados, Devisscher, and Martinez' (2015) study. The evaluation used cross-sectional agricultural and household surveys conducted in 2008 and 2009 to investigate the impact of land titles on agricultural production and agricultural investments. The study was based on a theory-based approach that hypothesized that land titles positively influenced access to credits and land investment rates (mechanization, cash crops, etc.) leading to an increase in agricultural production. It was assumed that households generally have credit restrictions and land titles are regarded as land tenure safeguards because financial institutions usually accept land titles as collateral for loans.

The evaluators argued that self-selection of the households that owned a title for their land did not make the sampling random. Therefore, they relied on a propensity score matching [7] to overcome the selection bias. The comparative statistical analyses of the titled households (treatment group) and the untitled ones (control group) showed no significant differences between the two groups in terms of agricultural production measured as the value of crop yields per hectare planted. The evaluation concluded that the theory of change allowed them to evaluate the impact of land title on agricultural production and investments.

While theory-based evaluation was gradually gaining ground, critics have alluded to the overrepresentation of donors' interests to the detriment of the local program communities by

[6] PTM "is a tool for within-case qualitative data analysis and refers to the 'systematic examination of diagnostic evidence selected and analysed in light of research questions and hypotheses posed by the investigator" (Bamanyaki & Holvoet, 2016, p.75).

[7] "A propensity score matching model is a two-stage procedure that addresses selection bias by comparing the outcomes of households with land titles with the outcomes of a selected group of households equivalent in all characteristic to those with lands without titles" (Hombrados, Devisscher & Martinez, 2015, p. 533).

focusing solely on accountability (Cousins, Whitmore, & Shulha, 2012). At the same time participatory research approaches that emerged in Asia, Latin America, and Africa seemed to provide alternative approaches in response to the quantitative models of study that could not answer the questions related to 'why' and 'how' the changes occurred (Cousins, Whitmore, & Shulha, 2012). Bamberger (2012) asserts: "when used in isolation, both quantitative and qualitative evaluation methods have strengths and weaknesses" (p. 3). Moreover, the conventional economic impact evaluation was not in line with complex development programs that required a diversity of methods and enhanced capacity (Lilja, Kristjanson, & Watts, 2010). The increasing complexity in the structure of development programs demands new and sophisticated approaches to evaluation as more and more emphasis is put on partnership and empowerment, thus the combination of the quantitative and qualitative approaches becomes necessary (Conlin & Stirrat, 2008). The debate about the most complete, valid, and rigorous method for impact evaluation led to the adoption of a mixed-methods approach.

Mixed-Methods Approach

The mixed-methods approach, made up of a combination of quantitative and qualitative methods (mostly participatory qualitative methods) and theoretical frameworks, focuses on broad and in-depth analysis and representation of the program under consideration without relying solely on statistics. It provides a more comprehensive knowledge of the issue under study, knowledge that is created from particularity to generality and contextual complexity with multiple perspectives (Greene, 2005). The experimental mixed-methods design relies mostly on quantitative data while the qualitative study is used as supplement. The participatory approach to impact evaluation uses qualitative research and involves evaluators working in collaboration with program participants. It is based on the concept that stakeholders must be included in some or all phases of the evaluation. The degree of participation of the stakeholders varies from program to program and from evaluator to evaluator, ranging from consultation to collaboration in decision making (Chambers, 1995; Cousins & Whitmore, 1998; Fetterman, 1994; Patton, 1994). This provides in-depth perspectives on the processes of change caused by the program (Mills & Gray, 2016). In order to validate the quantitative and qualitative data collected, the evaluator uses triangulation to extrapolate and compare data collected from various sources. Triangulation is a process of using various tools to examine the same phenomena (Mikkelsen, 2005).

Karuiki and Njuki's (2013) study is a very relevant example of a participatory impact evaluation combined with a quantitative method. The authors used participatory impact diagrams (PID) to evaluate gendered community perceptions of Arid Lands Resource Management Project (ALRMP) interventions in Kenya. Arid and semi-arid lands cover roughly 80% of the country and the pastoral populations of the arid and semi-arid lands are often affected by various climatic changes such as droughts, floods, and livestock diseases. The project's aim was to respond to the development priorities of Kenya's pastoral populations by reducing their vulnerability while reinforcing community-led income generation leading to food security and increased access to basic services like education, health care etc. It was a six-year project implemented in two phases; however, the discussion in this paper will focus only on the evaluation of the first phase of the project.

The PID is a participatory impact assessment tool that uses diagrams to evaluate both positive and negative impacts to development programs. The first implementation phase of the project was made up of a random sampling of 10 districts out of 22 districts. In each of the districts, two intervention sites were chosen and the evaluation was done in one intervention per site. For the qualitative assessment, the communities that received the intervention took part in the entire evaluation process. The impact diagrams were drawn on a large piece of paper or on the ground by the facilitator to represent the program. This allowed the participants (most of them illiterate) to identify the immediate and/or direct changes that occurred in the community and/or at the household level(s) as a result of the intervention. Negative changes were marked on one side and positive on the other, and a straight arrow was drawn from the program to the change. The PID from the qualitative approach was complemented by a quantitative survey in order to compare and contrast the data collected from various sources. The evaluation results revealed positive and negative changes including the identity of the participants and the number of participants who were affected by the changes.

Broegaard, Freeman, and Schwensen (2011) also used a mixed-methods approach to conduct an impact evaluation of socio-economic effects of improved transport infrastructure in Nicaragua.

The evaluation plan was based on an early analysis of the program's intervention logic, its implementation, the contexts in which it operated, and data availability. Related to the quantitative approach, a double-difference estimation [8] of quantifiable impact was used to determine the counterfactual [9] effect. The quantitative analysis was based on limited household-level data from the National Living Standards Survey in combination with data from National Census: 2001-2005. The quantitative data were supplemented by qualitative data collected from all regions that took part in the intervention. The qualitative data were made up of a total of thirty-nine (39) communities selected based on qualitative Participatory Rural Appraisal (PRA)[10] methods in the three regions of the interventions. Twenty-six (26) of the selected communities constituted the treatment groups and thirteen (13) were kept as control groups. To facilitate the comparison of quantitative and qualitative data, the same sampling criteria were used for both quantitative and qualitative data. The inclusion of qualitative methods provided the opportunity to assess the requirements for sustainability and the extension of the series of benefits which would not have been possible if the evaluators were to rely solely on quantitative data. Broegaard et al. concluded that the impact evaluation of the rural transport infrastructure has been very successful and proved that mixed-methods are very relevant in assessing the impact of complex interventions. They further assert that "the iterative process of contrasting and comparing results for analytical enhancement helped the evaluation to gain deeper understanding of quantitative results" (p.24).

[8] "Double difference is a technique that measures differences between the control group and the treatment group before and after the intervention" (NONIE, 2009, p.26)

[9] Counterfactual is the attempt to compare the change brought about by the intervention with what would have happened in the absence of the intervention (NOMIE, 2009, p.21).

[10] Participatory rural appraisal is an approach that includes the views, perspectives and knowledge of the rural stakeholders in the planning, management and evaluation of projects and programs. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Participatory_rural_appraisal on May 9, 2016.

Conclusion

In light of the discussions provided in the paper, I argue that impact evaluation requires a design that responds to contextual and situational aspects of development programs. This means that impact evaluations should be designed according to the requirements and limitations of the situation and the context. The best method is the one that will ask and effectively respond to the following questions as stated by Rogers (2009, p.218): (1) "What is the nature of the program? Why is an impact evaluation being done? What resources are available? And who would be using the evaluation results?" Once those questions have been answered, only then can the evaluator proceed with the selection of the most appropriate method(s) to use for the impact evaluation.

In this paper the definition of impact evaluation and the debates related to impact evaluation methods within the development context have been discussed. This was followed by the discussion of various methods used to conduct impact evaluation of programs in developing countries together with concrete examples of impact evaluations. The author's reflections and perspectives on impact evaluation and the current debates over the appropriate methods for impact evaluation in the development context have also been discussed.

The use of quantitative methods alone to conduct impact evaluation does not seem to answer the questions of 'why and how' changes occur. Different approaches have evolved including mixed-methods. There has been a considerable transformation from the call for improving impact evaluation where randomized controlled trials approach with causal attribution was advocated by the CGD. Now there is a growing list of tools and methods that can be used to conduct impact evaluations. Therefore, there is a need for program evaluators to tailor impact evaluation designs to the needs and constraints of the specific cases that will guide the appropriate methods selection process. Based on the reviewed literature, there are very limited examples of randomized control trials. Most of the impact evaluations within the development context used mixed-methods based on the combination of quantitative and participatory qualitative approaches including theory-based evaluations.

References

Bamanyaki, P.A., & Holvoet, N. (2016). Integrating theory-based evaluation and process tracing in the evaluation of civil society gender budget initiatives. *Evaluation, 22*(1), 72-90.

Bamberger, M. (2012). Introduction to mixed methods in impact evaluation. *Impact Evaluation Notes*, 3, 1-42. https://www.interaction.org/document/guidance-note-3-introduction-mixed-methods-impact-evaluation.

Bamberger, M. (2015). Innovations in the use of mixed-methods in real-world evaluations. *Journal of Development Effectiveness, 7*(3), 317-326.

Bamberger, M., Tarsilla, M., & Hesse-Biber, S. (2016). Why so many "rigorous" evaluations fail to identify unintended consequences of development programs: How mixed methods can contribute. *Evaluation and Program Planning, 55*, 155-162.

Beltramo, T., & Levine, D. I. (2013). The effect of solar ovens on fuel use, emissions and health: Results from a randomized controlled trial. *Journal of Development Effectiveness*, 5(2). 178-207.

Broegaard, E., Freeman, T., & Schwensen, C. (2011). Experience from a phased mixed-methods approach to impact evaluation of Danida support to rural transport infrastructure in Nicaragua. *Journal of Development Effectiveness, 3*(1), 9-27.

Cameron, D.B., Mishra, A., & Brown, A.N. (2016). The growth of impact evaluation for international development: How much have we learned? *Journal of Development Effectiveness, 8*(1), 1-21.

Cartwright, N. (2011). The art of medicine: A philosopher's views of the long road from RCTs to effectiveness. *The Lancet,* 377.

Carvalho, S., & White, H. (2004). Theory-based evaluation: The case of social funds. *American Journal of Evaluation*, *25*(2), 141-160.

Center for Global Development (2006). *When will we ever learn?* Washington DC: Center for Global Development.

Chambers, R. (1995). Poverty and livelihoods: Whose reality counts? *Environment and Urbanization, 7*(1), 173-204.

Collier, D. (2011). Understanding process tracing. *Political Science and Politics, 44*(4), 823-830.

Conlin, S., & Stirrat, R.L. (2008). Current challenges in development evaluation. *Evaluation, 14*(2), 193-208.

Coryn, C. L. S., Noakes, L. A., Westine, C. D., & Schröter, D. C., (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation*, *32*(2), 199-226.

Cousins, J.B., & Whitmore, E. (1998). Framing participatory evaluation. *New Directions for Evaluation, 80*, 5-23.

Cousins J. B., Whitmore, E., & Shulha, L. (2012). Arguments for a common set of principles for collaborative inquiry in evaluation. *American Journal of Evaluation, 34*(1), 7-22.

Fetterman, D. M. (1994). Empowerment evaluation. *Evaluation Practice, 15*(1), 1-15.

Greene, J.C. (2005). The generative potential of mixed methods inquiry. *International Journal of Research and Methods in Education*. *28*(2), 207-211.

Hearn, S., & Buffardi, A.L. (2016). Methods lab: What is impact? *Better Evaluation*. Retrieved from http://betterevaluation.org/resource/discussion-paper/what_is_impact.

Hombrados, J.G., Devisscher, M., & Martinez, M.H. (2015). The impact of land titling on agricultural investments in Tanzania: A theory-based approach. *Journal of Development Effectiveness, 7*(4), 530-544.

Karuiki, J. & Njuki, J. (2013). Using participatory impact diagrams to evaluate a community development project in Kenya. *Development in Practice, 23*(1), 90-106.

Lilja, N., Kristjanson, P., & Watts, J. (2010) Rethinking impact: Understanding the complexity of poverty and change overview. *Development in Practice, 20*(8), 917-931.

Mayne, J. (2012). Contribution analysis: Coming of age? *Evaluation, 18*(3), 270-280.

Mertens, D., & Tarsilla, M. (2014). *Mixed methods in evaluation*. In S. Hesse-Biber (Ed.) *Mixed Methods Handbook*, Oxford University Press.

Mikkelsen, B. (2005). *Methods for development work and research*. Thousand Oaks, CA: Sage Publications.

Mills, G. E., & Gray, L. R. (2016). *Educational research: Competencies for analysis and applications* (11th Ed.). Pearson Education Inc.

NONIE (2009). *Impact evaluations and development: Guidance on impact evaluation*. Retrieved from: www.worlbank.org/icg/nonie.

OECD/DAC. (2002). *Glossary of key terms in evaluation and results based managemen*t. OECD-DAC, Paris. Retrieved from http://www.oecd.org/dac/2754804.pdf.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice, 15*(3), 311-319.

Patton, M, Q. (2008a). *Utilization-focused evaluation*. (4th ed.). Thousand Oaks, CA: Sage Publications.

Patton, M. Q. (2008b). *State of the art in measuring development assistance*. Address of the World Bank Independent Evaluation Group Conference, 10 April, Washington, DC. Retrieved from http://www.worldbank.org/ieg/conference/results/patton.pdf.

Rogers, P. (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.

Stame, N. (2010). What doesn't work? Three failures, many answers. *Evaluation, 16*(4), 371-387.

Vaessen, J. (2010). Challenges in impact evaluation of development interventions: Opportunities and limitations for randomized experiments. *Institute of Development Policy and Management*. Discussion Paper/2010-01, 1-40.

Wang, P., Connor, A.L., Guo, E., Namboa, M., Chanda-Kapata, P., & Lambo, N. et al. (2016). Measuring the impact of non-monetary incentives on facility delivery in rural Zambia: a clustered randomized controlled trial. *Tropical Medicine and Institutional Health, 21*(4) 515-524.

Weiss, C. (1998). *Evaluation*. Upper Saddle River, N.J. Prentice Hall.

White, H., & Sabarwal, S. (2002). Quasi-experimental design and methods. *Methodological Briefs Impact Evaluation* No. 8. Retrieved from http://www.unicef-irc.org/publications/pdf/brief_8_quasi-experimental%20design_eng.pdf.

White, H. (2009). Theory-based impact evaluation: Principles and practice. *Journal of Development Effectiveness*, *1*(3), 271–284.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation*, *16*(2), 153–164.

World Bank, (2011). What is impact evaluation? Retrieved from: http://go.worldbank.org/2DHMCRFFT2.