Cite as: Pattyn, V. Eur J Dev Res (2019) 31: 174. https://doi.org/10.1057/s41287-019-00202-w

TOWARDS APPROPRIATE IMPACT EVALUATION METHODS

The choice of evaluation methods is one of the most plagued questions for evaluators (Szanyi, Azzam, & Galen, 2012). Especially in development evaluation, where interventions tend to be very complex, and multiple stakeholders hold competing interests (Holvoet et al., 2018), this question is pressing. While one can discern an emerging consensus among evaluation scholars that not only (quasi-) experimental evidence can lay monopoly claim to the production of the best effectiveness evidence (Stern et al., 2012), this idea is definitely not yet commonly shared among all evaluators, let alone among commissioners of impact evaluation studies. The article by Wendy Olsen presents a strong persuasive case for considering alternative impact evaluation methods that can help overcoming the shortcomings of Randomized Controlled Trials (RCTs). The question is, however, on which conditions one should opt for such alternative methods. Or to put it differently: on which conditions can it be "unwise" to resort to such methods in impact evaluations? The aim of this contribution is to bring some nuance into the methods debate, by drawing attention to the broader organizational and institutional context in which impact evaluations take place. The commentary revolves around the idea that the choice of a particular evaluation method will as much be affected by considerations of technical appropriateness as well as political appropriateness. With technical appropriateness, we refer to the ability of the chosen method to answer the impact evaluation question at stake. Political appropriateness in turn concerns the fit between the broader institutional setting and specific impact evaluation methods. Any assessment of the merit of particular evaluation methods should ideally consider both angles.

In line with Dr. Olsen's article (2019), the commentary particularly zooms into RCTs on the one hand, and case based approaches such as Qualitative Comparative Analysis (QCA) on the other hand. These methods represent two of the main design families that are discerned in impact studies, each relying on a different notion of causality (for an overview of all design families, see e.g. Stern et al., 2012). In a first section, we discuss the typical kind of evaluation questions that can be served per method. In a second part, we highlight some preconditions that should be fulfilled in the broader organizational and institutional setting for a successful implementation of alternative methods. While by no means providing a full picture of all criteria that should be met prior to engaging in an impact evaluation applying any of these methods (see e.g. Befani, B.; O'Donnell, 2016 for a more comprehensive checklist of criteria), we aim to give an indication of the importance of the 'politics of impact evaluations'.

Technical appropriateness

Project evaluations are intrinsically a political endeavor. Evaluation is intended to influence decision-making, whether decision makers are working in NGOs, governments, or other development organisations. Evaluation itself also has a political stance by making sometimes sensitive statements about issues such as the utility or relevance of development projects (Dahler-Larsen, 2012; C. H. Weiss, 1993). Against this background, development impact evaluations can serve multiple purposes for stakeholders. In the evaluation literature, a typical distinction is made between evaluations for accountability ('was the project money well spent?'), evaluations for implementation improvement ('how can we improve the effectiveness of the development project?') and evaluations for policy learning ('how can we explain project success or failure?') (Vedung, 1997). Although the typology is somehow analytical and seldom clear-cut in practice, the goal that the evaluation is meant to serve will largely determine how an evaluation will be used (Chelimsky & Shadish, 1997, p. 18). In developing an evaluation design, it is imperative to take this purpose into account. To be sure, evaluations can also be commissioned for symbolic reasons, for instance to justify decisions that were already taken. For the purposes of this contribution, though, we do not take such political-tactical goals into account.

As hinted at above, each of the evaluation purposes corresponds with a typical kind of evaluation question. When an accountability-oriented evaluation is at stake, stakeholders will mostly be interested in 'What works' to "speak truth to power" (Wildavsky, 1987). For evaluations that are set up for 'enlightenment' purposes (Weiss, 1977), more emphasis will be put on revealing 'What works for whom under which circumstances' as it is often coined (Pawson & Tilley, 1997). Importantly, counterfactual impact evaluations, such as RCTs, do serve a different kind of evaluation question than alternative case-based comparative methods. Whereas the former are primarily geared towards identifying 'whether' an intervention worked; case based methods will rather work towards understanding how an intervention works in context.

The different evaluation questions reflect different approaches to causality. Case-based approaches, such as Qualitative Comparative Analysis (QCA), are built on the idea of causal complexity in which context plays a central role. Underlying is the paradigm that social phenomena, including development interventions, have multiple and conjunctural causes. Whether a development project works, will not only depend on the characteristics of the project itself, but also on the context in which it is imbedded (Pattyn & Verweij, 2014). Development projects themselves, from such approach, are best to be conceived as 'contributory causes' which work as part of a causal package in combination with other factors (Stern et al., 2012). Instead of providing "monocausal explanations" (Sager & Andereggen, 2012, p. 6) comparative case based methods are designed to identify different combinations of conditions leading to a given outcome, i.e equifinality. The intervention is but one of such conditions, which will often not be sufficient in itself to trigger change. Case-based methods furthermore assume that causality is not necessarily symmetrical. The presence and absence of success of policy interventions can be explained by different causal 'recipes' (Ragin, 1987, 2000; Rihoux & Ragin, 2009).

By contrast, in experimental designs with their 'successionist' logic of causality (Pawson & Tilley, 1997), contextual variation is by design kept at a minimum between treatment and control group. The experimental approach resonates with a rational and positivist take on development interventions, and has the ambition to identify the average effect the intervention makes. By revealing which intervention works, the idea is to gradually come to the 'right' program theory that works best to achieve particular societal changes (van der Knaap, 2004). Whereas experimental designs are useful for accountability purposes, they are not the best approach for understanding which combinations of factors worked better under which circumstances of for different target groups (Befani, 2016, p. 17).

With impact evaluations being intrinsically political –in the broadest sense- it is important to have the purpose of the evaluation and the broader expectations of stakeholders sorted out from the outset of any development evaluation. On this basis, one can decide which method is technically most appropriate to use. No one method is superior under all circumstances.

Political appropriateness

Deciding on the purpose of an evaluation does not take place in a vacuum, but is particularly the result of the characteristics of the broader organizational and institutional context. Development cooperation is marked by often long donor-recipient chains, which implies that there is frequently a bias towards accountability oriented evaluations and the corresponding evaluation methods, at the expense of evaluations for policy learning and improvement. Yet, even if evaluation commissioners are explicitly interested in learning how an intervention works in context, this does not guarantee readiness for applying alternative impact assessment methods. To be clear, in reality, a plethora of evaluation questions is frequently put forward, both accountability as well as learning oriented. Yet also in these instances, it needs to be verified a priori whether the intervention is 'evaluable' with case-based methods, from a political-cultural stance. Case based methods come with important requirements for stakeholders in the field. These are sometimes at odds with what primary intended users are familiar with, particularly due to the longstanding dominance of methods as RCTs in many evaluation settings.

First, and as mentioned above, case based methods as QCA are built on assumptions of causal complexity. As a result, the corresponding summative lessons that are generated by this kind of alternative methods tend to be relatively complex, with the identification of multiple paths that lead towards program success or failure. A single condition proves seldom necessary and sufficient for societal change to occur. For evaluation commissioners and development practitioners complex causal paths can be puzzling (see also Pattyn, Molenveld, & Befani, 2017), as these will not allow drawing unambiguous conclusions that hold general relevance across contexts. While inherent to case-base approaches, it is up to the evaluator to verify upfront whether primary stakeholders are ready to walk this complexity path.

A successful application of a comparative case method as QCA furthermore requires case variability on conditions and outcomes. This implies that stakeholders should be willing to share information about different categories of cases, whether the intervention worked or not. Obtaining reliable information about 'failure' cases can be challenging, as aid recipients are not always keen to reveal such information (see Pattyn et al., 2017 for a more extensive discussion in this regard). To achieve an evaluation's learning-to-action potential to the fullest, in the sense of triple loop learning (Befani, 2016, p. 27), stakeholders themselves are preferably also involved in the selection of outcomes, conditions and in calibrating these. As such they can get a better understanding of what 'successful' impact means. It is clear that the realization of such ambitious learning goals will only be possible with the contribution of the key stakeholders involved.

The same precondition applies to the application of open retroductive methods: people should be willing to open the "can of worms" (Olsen, 2019). The iterative nature of case-based methods requires a long-term commitment of stakeholders involved in such evaluations. Own experience learns that this can be demanding, and expectation management is key in this regard. An evaluation from a case based perspective is in principle not concluded until all paths leading to a particular outcome are fully explained. It is particularly the goal of comparative case based methods to understand which combinations of conditions can account for success or failure, irrespective of whether such combination applies to typical or deviant cases. While RCTs can indeed be "slow" in generating impact evidence, one should be careful in concluding that alternative case-based methods are always much quicker. This can sometimes be true in formative terms. In summative respect, however, it can comparably take longer before a full understanding of "what works for whom in what circumstances" is acquired. After all, impact evaluations are in essence meant to reveal the "long-term effects produced by a development intervention'(OECD-DAC, 2002). Even if some interim effects can be brought to the attention of development practitioners during the evaluation process, an impact evaluation, *stricto sensu*, cannot yield final results before the longer term effects are known. No single evaluation method is exempt from this. Admittedly, summative results will often come in too late for decision makers. This is inherent to the nature of impact evaluations though, which often sits uneasily with policy decision timeframes. This being said, and considering the long timeframe of impact evaluations, it is definitely relevant to ask the question whether impact evaluations are always needed, and whether process or effectiveness evaluations are not sufficient for action purposes.

To conclude, and returning to the question leading this short commentary: On which conditions should one resort to case based methods? I fully concur with Olsen's call for working towards usable evidence that can bridge to action. The question is, however, what usable evidence involves, and whom it should serve. From the commentary, it should be clear that the answer will be contingent on the specific evaluation setting. The method(s) to be chosen should be consistent with the broader political purposes of the evaluation, and with the attitudes and values of the stakeholders involved. Where accountability motives are prioritized over learning purposes, it can be difficult to resort to alternative case based methods, even in a methodological pluralist setting.

References

Befani, Barbara; O'Donnell, M. (2016). Choosing appropriate evaluation methods tool. London: Bond. Retrieved from https://www.bond.org.uk/resources/evaluation-methods-tool

Befani, B. (2016). *Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA). Pathways To Change: Evaluating Development Interventions with QCA, Rapport till Expertgruppen för biståndsanalys (EBA)*. Retrieved from http://eba.se/wp-content/uploads/2016/07/QCA_BarbaraBefani-201605.pdf

Chelimsky, E., & Shadish, W. R. (1997). *Evaluation for the 21st century : a handbook*. Sage Publications.

Dahler-Larsen, P. (2012). *The evaluation society*. Stanford: Standford University Press.

Holvoet, N., Van Esbroeck, D., Inberg, L., Popelier, L., Peeters, B., & Verhofstadt, E. (2018). To evaluate or not: Evaluability study of 40 interventions of Belgian development cooperation. *Evaluation and Program Planning*, *67*, 189–199. http://doi.org/10.1016/j.evalprogplan.2017.12.005

OECD-DAC. (2002). *Glossary of Key Terms in Evaluation and Results Based Management*. Paris. Retrieved from http://www.oecd.org/dataoecd/29/21/2754804.pdf.

Pattyn, V., Molenveld, A., & Befani, B. (2017). Qualitative Comparative Analysis as an Evaluation Tool. *American Journal of Evaluation*, 109821401771050. http://doi.org/10.1177/1098214017710502

Pattyn, V., & Verweij, S. (2014). Beleidsevaluaties tussen methode en praktijk: Naar een meer realistische evaluatiebenadering. *Burger, Bestuur En Beleid. Tijdschrift Voor Bestuurskunde En Bestuursrecht*, *8*(4), 260–267.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.

Ragin, C. (1987). *The comparative method. Moving beyond qualitative and quantitative strategies*. London: University of California Press.

Ragin, C. (2000). *Fuzzy set social science*. Chicago: University of Chicago Press.

Rihoux, B., & Ragin, C. (2009). *Configurational comparative methods. Qualititative comparative analysis (QCA) and related techniques.* Thousand Oaks and London: Sage.

Sager, F., & Andereggen, C. (2012). Dealing With Complex Causality in Realist Synthesis: The Promise of Qualitative Comparative Analysis. *American Journal of Evaluation*, *33*(1), 60–78. http://doi.org/10.1177/1098214011411574

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the Range of Designs and Methods for Impact Evaluations. *Department for International Development*, (February 2011), 1–127. Retrieved from http://www.dfid.gov.uk/Documents/publications1/design-method-impact-eval.pdf

Szanyi, M., Azzam, T., & Galen, M. (2012). Research on evaluation: A needs assessment. *Canadian Journal of Program Evaluation*, *27*(1), 39–64.

van der Knaap, P. (2004). Theory-Based Evaluation and Learning: Possibilities and Challenges. *Evaluation*, *10*(1), 16–34. http://doi.org/10.1177/1356389004042328

Vedung, E. (1997). *Public policy and program evaluation*. Transaction Publishers.

Weiss, C. H. (1977). Research for Policy's Sake: The Enlightenment Function of Social Research. *Policy Analysis*, *3*, 531–545. http://doi.org/10.2307/42783234

Weiss, C. H. (1993). Where Politics and Evaluation Research Meet. *American Journal of Evaluation*, *14*(1), 93–106. http://doi.org/10.1177/109821409301400119

Wildavsky, A. (1987). *Speaking Truth to Power: Art and Craft of Policy Analysis*. London: Routledge. Retrieved from https://www.taylorfrancis.com/books/9781351488471