Special Debate Section

What Can We Learn from Impact Evaluations?

Robert Lensink[a,b]

[a]Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands.
[b]Wageningen University, Social Sciences, Wageningen, The Netherlands.
E-mail: b.w.lensink@gmail.com

European Journal of Development Research (2014) **26,** 12–17. doi:10.1057/ejdr.2013.43

Introduction

In the past decade, non-governmental organizations have been increasingly pressed to demonstrate whether their projects and programs are effective and contribute to enhancing the well-being of their beneficiaries. Although monitoring and evaluation have always been part of aid policies, most traditional evaluation methods come in for criticism because of their lack of rigor, selection biases and inability to address cause-and-effect questions. A recent surge in impact evaluations represents a response to the growing demands of policy makers that they be better informed about the impact of aid policies in general and specific aid projects in particular. Such impact evaluations must be clearly differentiated from evaluations in general: Whereas traditional evaluations address questions about the design or implementation of a project/program, impact evaluations are structured around attribution questions, namely, whether the change in outcomes is caused by the intervention. The main goal of an impact evaluation is therefore to measure the difference in an outcome, in a way that attributes the difference to the focal program and only that program. By learning more about their impacts, it may be possible to design better programs and products.

In this article, I argue that impact evaluations, if properly designed, are especially beneficial during the pilot stage of a project, to determine whether and in which conditions an intervention is likely to work. Furthermore, impact evaluations can enable tests of the relevance of different theories. The scope of rigorous impact evaluations of ongoing aid projects implemented in the past is rather limited though, which suggests the need for a serious evaluability assessment, even before deciding whether it is relevant to start an expensive impact evaluation. Moreover, most existing impact evaluations are stand-alone efforts, making it risky to base public aid policies on this limited evidence.

The Importance of Valid Comparison Groups

Formally, measures of program impact require comparisons of the same individual (or group of individuals) with and without aid programs at the same point in time. Obviously, such an evaluation is not possible; every individual is unique and has only one existence, and thus we can never observe the same individual with and without the program at the same point in time. Evaluators thus confront the problem of a missing counterfactual: We do not know what would have happened without the program.

Impact evaluations seek to uncover the causal effects of the intervention. Depending on the situation, they use different, mainly quantitative evaluation methodologies, though the list of research designs is quite extensive, including experimental approaches such as randomized controlled trials (RCTs), as well as quasi-experimental approaches such as instrumental variable (IV), regression discontinuity (RD), difference-in-difference (DID) and propensity-score matching (PSM) approaches, along with their combinations. The evaluation methods differ in several respects, though they all, in one way or another, try to deal with the problem of missing counterfactuals. That is, they try to assess what would have happened without the intervention, by defining a comparison or control group.

Comparison groups are crucial for impact evaluations, yet a valid impact evaluation cannot be assumed simply from a comparison of members of a program with non-members, because the two groups likely differ for reasons unrelated to the aid project. For example, a direct comparison of members and non-members probably suffers from program placement bias and individual selection bias, because most social interventions are targeted, and individual participation is voluntary, such that it depends on individual characteristics. The key challenge of a high-quality impact evaluation is therefore to find a control group that (a) in the absence of the intervention is identical to the treatment group, (b) reacts to the intervention similarly to the way the treatment group reacted, and (c) is exposed to the same set of external interventions as the treatment group (Gertler *et al*, 2011). If these three conditions hold, different outcomes for the two groups can be attributed to the aid intervention.

Scope and Relevance of Random Experiments

Analogous to drug testing, randomized experiments, which determine treatment and control groups by randomized assignment of the intervention, may provide a meaningful methodology for conducting valid impact evaluations; in certain conditions, this methodology can produce equivalent control and treatment groups. The RCT approach in development economics is the brainchild of economists such as Esther Duflo and her colleagues from Innovations for Poverty Action and the Abdul Latif Jameel Poverty Action Lab at MIT, who developed a methodology to conduct small experiments related to development aid anywhere in the world. It involves choosing communities or groups at random for a certain intervention. The effect of the program can then be measured by comparing post-intervention averages for the different outcome variables across the treatment and control groups. In turn, RCTs are becoming increasingly popular, especially as a method to evaluate existing projects. Furthermore, RCTs appear extremely valid for policy makers, and as various cases have shown, RCTs already influence policy making: A famous RCT of deworming programs in Kenya (Miguel and Kremer, 2004) offers a good example.

In principle, randomized experiments provide the best opportunity to rigorously examine causality questions. In addition, a properly designed RCT allows for impact evaluations conducted with post-treatment data. Yet despite their theoretical status as the 'gold standard' of impact evaluations, in practice, the conditions for *ideal* RCTs almost never hold, such that they require additional econometric techniques and statistics to control for remaining selection biases, differential treatment effects and so forth (Deaton, 2010). This point is not to suggest that RCTs are invalid; on the contrary, the biases due to less than ideal RCTs tend to be even more pronounced with other evaluation techniques. In principle, even RCTs conducted in non-ideal conditions provide the best opportunities to control for selection biases and ensure the internal validity of the impact evaluation. However, the impact evaluation outcomes of properly conducted RCTs still demand careful analysis, and perhaps validation by additional post-treatment qualitative and quantitative

research. To test how policy interventions affect beneficiaries, it is also important to obtain information about changes in their attitudes, such as risk and time preferences. Few impact evaluations pay much attention to this type of variable though, because as Harrison (2014) argues in this issue, impact evaluations normally focus on observables. This critique also extends beyond RCTs; it holds for almost all commonly applied quantitative and qualitative impact evaluation techniques. This caution further implies that the quality of impact evaluations might improve considerably if behavioral games were added to the impact evaluation protocol, possibly in combination with RCTs, to gather more evidence on several hard-to-measure variables.

The Need for Pre-Intervention Data

Which evaluation alternatives are available if RCTs are not possible? The main challenge, in my view, is the question of how to control for selection biases. It requires, at the very least, available data about non-treated individuals or groups, to support the construction of valid comparison groups. If these data are available, the use of IV methods is preferable, as long as appropriate instruments are available to identify the model, which generally requires exclusion restrictions. That is, valid instruments must satisfy orthogonality conditions, have a high correlation with the intervention and be properly excluded from the model, so that they only affect the outcome variables indirectly. In practice, appropriate instruments are almost never available, and therefore IV methods are rarely feasible for practical impact evaluations. In the absence of proper instruments, a PSM method may be applied, though this matching must apply to observed characteristics that are not affected by the intervention, and thus when it relies on post-treatment data, this method incurs great risk. A rigorous impact evaluation requires matching at the baseline, which is possible only if pre-intervention data are available. Another enormous advantage of having baseline information is that DID methods can be applied, possibly in combination with PSM, to control for selection biases that result from the time-invariant unobservable characteristics that affect participation. In the absence of randomly assigned control groups, proper impact evaluations can be conducted only if pre- and post-treatment data are available for the treatment individuals and groups and a comparison group. Even in the case of a randomized experiment, baseline information often turns out to be very important.

The Relevance of Impact Evaluations of Ongoing Aid Projects

How relevant is an impact evaluation of an aid project rolled out some time ago? In other words, to what extent is it possible to address causality issues using a *retrospective* evaluation? To answer this question, it is helpful to recall the conditions for a rigorous impact evaluation: valid comparison groups, preferably determined by randomized assignments, with accessible post- and pre-treatment data.

For many ongoing aid projects, the determination of valid comparison groups remains problematic, or even impossible. For some types of interventions, it turns out to be very difficult to find comparison groups that are not contaminated by similar types of interventions. For example, considering microfinance interventions, especially in some Asian countries, nearly everyone has access to some kind of microcredit or a similar financial product. Many aid projects contain interventions that are open to everybody, such as awareness or sensitization campaigns. In these situations, finding comparison groups is possible but obviously very problematic. Traditionally, evaluations of aid projects have been conducted in the absence of a comparison

group, using comparisons of the individuals or groups of individuals before and after the program. This so-called reflexive impact evaluation method cannot address claims of causality with rigor though, because no one can identify which changes were caused by the program and which changes resulted from other factors that also changed, around the time of treatment.

Even more problematic is the fact that, for many ongoing aid projects, pre-intervention baseline data are simply not available. Some baseline data might exist for the beneficiaries of the project, but almost never do they appear for a comparison group. Thus, it might be possible to proxy baseline information using recall procedures, which is clearly sub-optimal. Alternatively, the evaluation could use post-treatment data only. In both cases, it would be very difficult, if not impossible, to address causality questions.

The impact evaluations of aid projects implemented in the past become even more difficult when we recognize that, in practice, almost all aid projects contain multiple interventions. In these cases, it may still be possible to examine the extent to which entire programs cause changes in well-being. However, a theory-driven impact evaluation that precisely maps out and tests how and through which channels a certain intervention affects outcomes would be impossible for multiple interventions that cannot be disentangled.

For these reasons, the scope for rigorously addressing causality problems by evaluating ongoing aid projects is limited – which is not the same as assuming that the evaluation of these projects is useless. Instead, evaluations help answer questions about what is taking place, whether intended interventions have been conducted, and whether objectives have been accomplished. They simply may not be relevant for addressing cause-and-effect questions. Moreover, rigorous impact evaluations are often very costly, such that many ongoing aid projects and aid programs might not justify a broadly defined, high-quality impact evaluation. A more cost-effective, 'light' evaluation is probably preferable. In this issue, Guijt and Roche (2014; see in particular their Box 5) and Camfield and Duvendack (2014) argue that cost-effective alternatives are available for assessing causal inference, even when a valid counterfactual is not available or pre-intervention baseline data are missing. Camfield and Duvendack (2014) refer to the 'life histories' and 'process tracing' methods. I find relevance in these approaches, mainly in terms of a contribution analysis, but wonder whether they can really provide rigorous evidence about the attribution question. I also agree with Guijt and Roche (2014) that different policy-relevant impact analyses have different purposes, and rigor does not automatically imply relevance. To make an impact analysis policy relevant, we need more. Still, in my view, a rigorous impact evaluation requires a rigorous assessment of causality to start. Therefore, I call for detailed, serious analyses of the type of aid projects that warrant rigorous impact analyses. That is, researchers should conduct serious evaluability assessments, even before beginning the impact evaluation of any ongoing aid project. Obviously, as Guijt and Roche (2014) argue, transparency is an important condition for improving the relevance of impact evaluations. Many modern experimental impact studies therefore include so-called pre-analysis plans, to improve their transparency and avoid data mining (for example, Casey *et al*, 2012). Prior to the actual impact analysis, such plans can describe the intervention that needs to be analyzed, provide a relevant change theory and offer a detailed description of the intended data analysis.

When Are Impact Evaluations Relevant?

The scope for impact analyses of ongoing aid projects is limited, but they are likely particularly important in the pilot stage of a new project. Before rolling out a project across the eligible population, a rigorous impact analysis can provide the necessary information regarding whether

and in which conditions a clearly defined intervention works. An RCT is then most appropriate, especially in terms of testing the validity of conflicting theories.

The effectiveness of an aid project may also benefit greatly from field experiments if, before a development project gets implemented on a large scale, a randomized experiment tests whether it is likely to work. This point suggests the relevance of making impact evaluations an integral part of the aid project, right from the start. The implementation of independent, qualified impact evaluations during the pilot stage of a project could be critical. The credibility and usefulness of an impact evaluation depends on the quality of the evaluators and, perhaps even more important, on the willingness of project officials and evaluators (researchers) to work together and openly discuss project specificities. A close relationship between researchers and project managers can also encourage the development and testing of a theory of change (TOC) for the project. Impact evaluations must start by developing a TOC, usually by using a results chain. The TOC maps the entire process from inputs to outputs and explains whether and through which channels the intervention is likely to affect outcomes. Most important, a TOC obliges project managers to define project aims clearly and think carefully about the causal effects of the interventions.

Moreover, as correctly argued by Picciotto (2014) in this volume, a policy-relevant evaluation should examine not only whether the intervention works but also how relevant and efficient that intervention has been. Many impact evaluations focus on efficacy question, paying (too) little attention to efficiency. But determining efficiency, or conducting a less demanding cost-effectiveness analysis, is not obvious; it requires cost-benefit analyses, as argued by both Harrison (2014) and White (2014) in this issue. Moreover, a cost-effectiveness evaluation can be conducted only if we know the impact of the intervention. That is, a rigorous impact analysis is a prerequisite of a rigorous cost-effectiveness analysis.

Finally …

Impact evaluations during the pilot stage of a project may provide important information about whether the project should be undertaken. They should be used as preconditions for financing a project. However, even if an impact evaluation suggests positive impacts of a particular intervention during a pilot stage, it remains very risky to use these results as guidance for national aid policies. A rigorous impact evaluation controls for program placement biases and selection biases. That is, a rigorous impact evaluation improves the internal validity of the evaluation and provides important information about the relevance of a particular type of intervention. But because the external validity of the evaluation may be very low, there is no guarantee that impact evaluation results related to a certain project, even if the most rigorous impact evaluation technique has been used, will hold in other settings in the same country, let alone in other countries. The problem of poor external validity grows even more pronounced when we consider that, until recently, even the most rigorous impact evaluations were conducted as stand-alone efforts. Using stand-alone exercises, not validated by replication studies in other settings, impact analyses can never truly guide national aid policies. As White (2014) argues in this issue, synthetic reviews or statistical meta-analyses may be relevant to test the external relevance of different types of interventions. But precisely because there are so few studies dealing with similar types of interventions, the relevance of statistical meta-analyses in the context of impact analyses necessarily remains limited until more replication studies have been conducted. None of these cautions should be taken to mean that I don't see a future role for impact analyses, in terms of influencing national aid policies. On the contrary; I fully support White's (2014) plea for more rigorous impact analyses, which can create a situation in which aid policies can reasonably be based on interventions with high external validity. Yet not all aid projects qualify for

such a rigorous impact evaluation. We need to be selective precisely because there are limited resources available for impact analyses. Eventually, in my view, we will learn the most about which interventions work or fail by concentrating on the type of interventions that qualify for rigorous impact evaluations.

References

Casey, K., Glennerster, R. and Miguel, E. (2012) Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *Quarterly Journal of Economics* 127(4): 1755–1812.

Camfield, L. and Duvendack, M. (2014) Impact evaluation: Are we 'off the gold standard'? *European Journal of Development Research* 26(1): 1–11.

Deaton, A. (2010) Instruments, randomization, and learning about development. *Journal of Economic Literature* 48(2): 424–455.

Gertler, P., Martinez, S., Premand, P., Rawlings, L. and Vermeersch, C. (2011) *Impact Evaluation in Practice,* World Bank, http://www.worldbank.org/pdt.

Guijt, I. and Roche, C. (2014) Does impact evaluation in development matter? Well, it depends what its for!. *European Journal of Development Research* 26(1): 46–54.

Harrison, G.W. (2014) Impact evaluation and welfare evaluation. *European Journal of Development Research* 26(1): 39–45.

Miguel, E. and Kremer, M. (2004) Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72(1): 159–217.

Picciotto, R. (2014) Is impact evaluation evaluation? *European Journal of Development Research* 26(1): 18–25.

White, H. (2014) Current challenges in impact evaluation. *European Journal of Development Research* 26 (1): 18–30.