FONDATION POUR LES ÉTUDES ET RECHERCHES SUR LE DÉVELOPPEMENT INTERNATIONAL

Working Paper 51 October 2012 Development Policies

How Can Development NGOs Be Evaluated?

Jan Willem Gunning

Jan Willem Gunning is Professor of development economics and director of the Amsterdam Institute for International Development (AIID). He also has been a staff member of the World Bank and Professor at the University of Oxford where he directed the Centre for the Study of African Economies (CSAE). His research interests include poverty dynamics, impact evaluation, and the effect of risk on growth in rural societies.

1. Introduction

In the past decade there has been a remarkable surge of interest in using impact evaluation to establish the effectiveness of development interventions. The conventional evaluation methods used by development consultants typically rely on simplistic before-after or with-without comparisons. Impact evaluation, by contrast, involves comparing actual outcomes with a formal counterfactual. Any differences between the two can be attributed to the intervention if, and only if, the counterfactual is credible. A conventional before-after comparison fails this test since the "before" situation is obviously not a credible counterfactual: outcomes could have changed over time for reasons unrelated to the intervention. Randomized controlled trials (RCTs) are usually considered the preferred design for impact evaluation, but there are also regression-based techniques such as regression discontinuity designs and regressions in first differences (double differencing). . . . /. . .

LA FERDI EST UNE FONDATION RECONNUE D'UTILITÉ PUBLIQUE.
ELLE MET EN ŒUVRE AVEC L'IDDRI L'INITIATIVE POUR LE DÉVELOPPEMENT ET LA GOUVERNANCE MONDIALE (IDGM).
ELLE COORDONNE LE LABEX IDGM+ QUI L'ASSOCIE AU CERDI ET À L'IDDRI.
CETTE PUBLICATION A BÉNÉFICIÉ D'UNE AIDE DE L'ÉTAT FRANÇAIS GÉRÉE PAR L'ANR AU TITRE DU PROGRAMME «INVESTISSEMENTS D'AVENIR» PORTANT LA RÉFÉRENCE «ANR-10-LABX-14-01».

Impact evaluation is used by many donor agencies to assess the effectiveness of aid-supported interventions in developing countries. In addition, increasingly governments in these countries evaluate their own programs. Evaluation of development NGOs has lagged behind although one of the best known evaluation articles (usually referred to as the "de-worming paper", Miguel and Kremer, 2004) involved a Dutch NGO active in primary education in Kenya. Superficially, whether a program or project is run by an NGO or by a government agency is irrelevant for the design of an evaluation. Indeed, in many cases standard impact evaluation techniques can be used to evaluate the activities of NGOs.

NGOs often resist impact evaluation. Some of this is based on irrational opposition to rigour and quantitative analysis. But opposition is sometimes based on the claim that what NGOs do makes them fundamentally different and unsuitable for standard evaluation methodologies such as RCTs. While this claim is usually poorly articulated it deserves to be taken seriously. In this paper we consider two ways in which NGOs can indeed be "different" and the implications for evaluation design.

First, many NGOs try to achieve their goals indirectly, notably through advocacy. For example, an NGO may aim to reduce infant mortality or improve the quality of teaching. However, they do this indirectly, through activities aimed at achieving changes in policies, rather than directly, e.g. by setting up a project to train teachers. Clearly, the implicit theory of change involves two steps: advocacy succeeds in changing policies and policy changes lead to the intended development outcomes. When the policies are local (e.g. nursing practices at health clinics which affect child mortality) then a standard RCT approach can be used: randomization would then be over health clinics and advocacy would be limited to those in the treatment group.[1] However, when the policies cover wide areas there is little scope for RCTs and in the case of national policies there is none. This is a serious problem for the evaluation of many NGOs.

Secondly, NGOs are often highly decentralized. Decisions are delegated externally to partner organizations and internally to field offices and local staff. For example, an NGO may offer sanitation training to communities, selecting amongst eligible communities those where staff with local knowledge expect the effect to be greatest. Standard evaluation designs are then useless: they would necessarily assign the treatment (or at least the intention to treat) to communities in a way which differs from the way the allocation would actually be made since the latter is based on private information of local staff. This problem is not exclusive to NGOs but it is particularly likely to be relevant in evaluations of NGOs because of the nature of their organizations.

These problems and their possible solutions are discussed in turn in sections 2 and 3. Section 4 concludes.

1. In The Netherlands the government has commissioned a massive evaluation, involving a sample of all government supported activities of Dutch development NGOs. The study design involves many examples where RCTs are used with randomization over locations. See, for example, the description of the Indonesia part of this evaluation: http://www.aiid.org/.

2. Advocacy: Achieving Change Indirectly

An example of an NGO which tries to achieve change through advocacy is the East African NGO Twaweza, based in Dar es Salaam.[2] Twaweza aims at enabling the poor to exercise agency and thereby to gain access to basic services. It does so in part by fuelling debate and heightening aspirations. Twaweza's approach is based on a sophisticated theory of change. Twaweza is unusual in that it does not directly aim at a goal like better education or health services but tries to achieve such outcomes indirectly. Many of its activities are aimed at creating awareness of a particular problem at the national level. Twaweza may draw attention to a problem by organizing a workshop, getting newspaper coverage of the issue, or by talking to political or religious leaders. It is hoped that such activities will result in information on the topic reaching villages and there triggering discussions. Twaweza expects that this in turn will prompt "agency" (or "public action") and eventually improved outcomes.

Figure 1 summarizes in a simplified way in which Twaweza initiatives aimed at national organizations and institutions may eventually affect outcomes at the local level. The ongoing evaluation (2012-2015) is described as follows in AIID (2011) in terms of this Figure:

> "Twaweza activities are pictured around the circled (1) in the graph. They are aimed at various organizations and institutions. In turn these organizations call on local organizations at the village level, leading to local interventions and activities. Obviously, organizations both national and local, can also be influenced by initiatives not originating from Twaweza. Local interventions (pictured around the circled 2) can lead to public action at the local level and subsequently to outcomes in terms of service delivery or household indicators on health, education and access to safe water facilities. The evaluation design's quantitative part focuses on the elements around circled (2): how local initiatives trigger public action and how public action affects outcomes. However, local initiatives need not have been prompted by Twaweza activities: they could have many origins. Establishing the relationships around (2) would however provide evidence for a central piece of Twaweza's theory of change: that local activities and agency can lead to improved public service delivery.
>
> Assuming that it can be shown that relationship (2) works as theorized it remains to establish that Twaweza, although it may not be responsible for all local initiatives, can and does trigger at least some of them. This is the part pictured around circled (1). In the evaluation we will follow the links under (1) qualitatively by a number of case studies and quantitatively by investigating whether there exists an 'information echo' of Twaweza activities in national and local media. This requires qualitative research similar to expenditure tracking: the researcher attempts to trace the effect of the Twaweza intervention through the organizations Twaweza has talked to, down to the local level. This can obviously not be done exhaustively or very accurately, but it is important to establish plausibility of a channel from the intervention to the local event. "

2. See http://www.twaweza.org/. The Twaweza example in this section is based on AIID (2011).

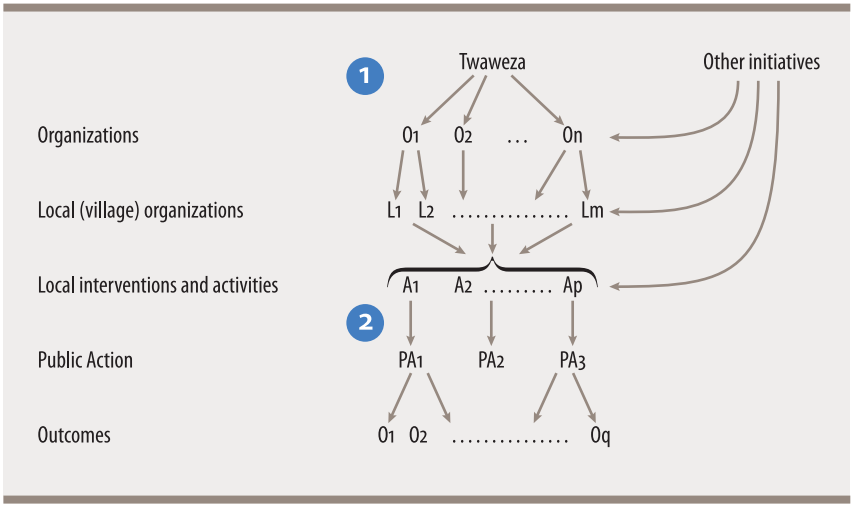

Figure 1. Simplified intervention logic for Twaweza
Source: AIID (2011)

This evaluation uses two types of data. First, household surveys are conducted at two points in time, a baseline in 2011 and a final survey in 2014. These surveys record household characteristics, information at the household level, participation in collective action and health and education outcomes. The sample households live in 250 randomly drawn locations throughout Tanzania.

Secondly, during the three years between the two surveys data are collected through high frequency monitoring. Carefully selected informants in each of the 250 villages are interviewed every three weeks by cell phone. The phone interviews cover a wide range of topics, notably (a) what information reaches the village? For example, the informant may report that people have heard on the radio that the quality of schooling in Tanzania is poor and that this has led to (or reinforced) dissatisfaction about local schooling (b) whether such dissatisfaction led to public action, e.g. through pressure on teachers or complaints to politicians.[3]

Step 2 in Figure 1 can be tested formally. Regressions can be used to explain changes over time in relevant outcomes such as learning outcomes in terms of local public actions (recorded in the phone interviews), controlling for changes in household and community differences. This will establish whether public action is effective in improving outcomes.

An obvious objection is that public action is clearly endogenous in such a regression: communities may differ systematically in the information they receive and in how they respond to that information. In other words: public action has not been randomly assigned across villages. This concern is misplaced: this selection effect is an integral part of the program that is being evaluated. One should therefore not try to correct for it.[4] For example, if Twaweza has adopted communication strategies which imply that information is more (less) likely to arrive in those villages where it triggers effective collective action, then the program is to that extent more (less) effective than a program which would lead to a random distribution of information over villages. Clearly, the evaluation should take this into account.

3. The network of informants is described by Twaweza as a listening device, a unique way of collecting information on what people know, care about and do for a nationally representative sample. It may be seen as up-scaling of the approach of a traditional anthropologist to a large and nationally representative sample.
4. On this point see Elbers and Gunning (2012).

Step 1 cannot be analyzed in a similar way. If the high frequency monitoring does not pick up any echo of Twaweza's national activities that would be convincing evidence: Twaweza's theory of change would be rejected. But the case is not symmetric: if there are lively village discussions on the poor quality of education this might be a result of national Twaweza campaigns on this subject but also could also reflect activities of other NGOs or of the government. Here the best one can do is use qualitative analysis of what Twaweza did at what time and what others did to build a plausible case that the interventions can (or cannot) be attributed to Twaweza.

The example illustrates how NGO activities may not be suitable for RCTs. However, it also shows that a substantial part of the theory of change (on the effect of information on collective action and the effect of the latter on development outcomes) does lend itself to rigorous analysis, contrary to what is often suggested in similar contexts.

3. Imperfect Control [5]

In medical research the distinction between efficacy and effectiveness is well established. Efficacy refers to the effect of a treatment in tightly controlled conditions in a laboratory or in a (randomized) clinical trial. A successful clinical trial is seen as "proof of principle". Effectiveness denotes the effect of the treatment in practice, under normal conditions, that is typically with imperfect control. Efficacy is a necessary but not a sufficient condition for effectiveness. For example, HIV/AIDS drugs are much less effective in Africa than one would expect on the basis of clinical trials since patients often share drugs with others.

The distinction is not as well established in development. RCTs, designed to establish efficacy ("proof of principle") are used as if their results establish the effectiveness of actual policies. In this section we consider a plausible situation in which a standard RCT evaluation of an NGO activity is inappropriate.

To understand the issue it is useful to consider the following model:

$$y_{it} = \alpha_t + \beta_i P_{it} + \gamma X_{it} + \eta_i + \varepsilon_{it} \qquad t = 0,1 \qquad (1)$$

where y measures an outcome of interest, P is a vector of the interventions to be evaluated and X a vector of observed controls; η_i represents the combined effects of unobserved characteristics (assumed to be time invariant) and ε_{it} is the error term, assumed to be independent over time; $t = 0,1$ is the time of measurement; and $i = 1,\ldots, n$ denotes the unit of observation (*e.g.* households or locations). P can be either actual treatment or intention to treat. Assume that the interventions and control variables are uncorrelated with the error process:

$$X_{i1}, X_{i0}, P_{i1}, P_{i0} \perp \varepsilon_{i1}, \varepsilon_{i0}.$$

5. This section draws heavily on Elbers and Gunning (2012).

and that P and X are independent: [6]

$$X_{i1}\,,\, X_{i0} \perp P_{i1},\, P_{i0}\,.$$

However, we do not assume

$$P_{i1},\, P_{i0} \perp \eta_i$$

so that assignment may be correlated with unobserved characteristics.

The textbook evaluation is a special case of (1): (a) P is a variable rather than a vector, (b) P takes only two values (1 for treatment, 0 for control), and (c) there is no treatment heterogeneity so that

$$\beta_i = \beta$$

for all i. Clearly, β is then the parameter of interest and it can be identified by an RCT or, equivalently by a regression in terms of differences (which eliminates the endogeneity of P resulting from its correlation with η):

$$\Delta y_i = \beta \Delta P_i + \gamma \Delta X_i + \Delta \varepsilon_i \qquad (2)$$

However, this special case is indeed very special and may not be relevant for an NGO evaluation. Notably, problems can arise if there is treatment heterogeneity so that (2) becomes

$$\Delta y_i = \beta_i \Delta P_i + \gamma \Delta X_i + \Delta \varepsilon_i \qquad (3)$$

and if P is no longer a binary variable and β_i and ΔP_i are *not* independent. For example, an NGO program may be implemented by program officers who can use their discretion in choosing ΔP_i and who have some private information on β_i. They might then focus the intervention on the more promising households or communities ("targeting on the gain" in the terminology of Heckman), resulting in a correlation of β_i and ΔP_i. Such situations are likely to be important in practice: the effectiveness of interventions will typically differ across beneficiaries and it is quite plausible that NGOs staff with local expertise will have some knowledge of this treatment heterogeneity and use it in deciding where (and how intensively) to apply the intervention. In that sense an NGO policy maker has limited control: assignment decisions are often delegated to lower level staff ("program officers") precisely because these have relevant private information.

For the evaluator, however, this situation presents a very serious challenge. An RCT would impose assignments (or at least intentions to treat) ΔP_i randomly over beneficiaries and would thereby produce an unbiased estimate of the population average $E\beta_i$. This is, however, irrelevant as a measure of the program's effectiveness since in actual practice assignment is not random. An NGO's resistance to an RCT evaluation would in this case be entirely justified. Similarly, if the evaluator uses observational data (possibly from NGO records) instead of experimental data then a double difference regression would estimate

$$\Delta y_i = \beta \Delta P_i + \gamma \Delta X_i + \omega_i \qquad (4)$$

6. This can be relaxed: Elbers and Gunning (2012), section 3.

where

$$\beta = E(\beta_i | \Delta X_i, \Delta P_i)$$

and

$$\omega_i = \Delta\varepsilon_i + (\beta_i - \beta)\Delta P_i$$

and this suffers from endogeneity as a result of the correlation between β_i and ΔP_i.

Elbers and Gunning (2012) propose a solution. They approximate the expectation of β_i linearly:

$$E(\beta_i | \Delta X_i, \Delta P_i) \approx \delta_0 + \delta_1 \Delta X_i + \delta_2 \Delta P_i$$

so that

$$\Delta y_i = \alpha_1 \Delta X_i + \alpha_2 \Delta P_i + \alpha_3 \Delta X_i \otimes \Delta P_i + \alpha_4 \Delta P_i \otimes \Delta P_i + \omega_i \qquad (5)$$

where $\alpha_2 \Delta P_i + \alpha_3 \Delta X_i \otimes \Delta P_i + \alpha_4 \Delta P_i \otimes \Delta P_i$ is the approximation of $T_i = E(\beta_i \Delta P_i | \Delta X_i, \Delta P_i)$.

Equation (5) does not suffer from endogeneity so that OLS will produce unbiased estimates. The estimated coefficients can then be used to estimate T_i as

$$\hat{T}_i = \hat{\alpha}_2 \Delta P_i + \hat{\alpha}_3 \Delta X_i \otimes \Delta P_i + \hat{\alpha}_4 \Delta P_i \otimes \Delta P_i.$$

If the data are from a random sample of beneficiaries then the total program effect (TPE) can be estimated as the average of $\hat{T}_i$ in the sample:

$$\hat{TPE} = \frac{1}{n} \sum_i \hat{T}_i = \hat{\alpha}_2 \overline{\Delta P_i} + \hat{\alpha}_3 \overline{\Delta X_i \otimes \Delta P_i} + \hat{\alpha}_4 \overline{\Delta P_i \otimes \Delta P_i} \qquad (6)$$

where bars denote sample averages.

In this situation a standard RCT would, as we have seen, not achieve internal validity. (This is worth stressing since it is commonly assumed that the internal validity of RCTs is beyond question.) However, the RCT could be designed differently: by randomizing over program officers rather than beneficiaries. Program officers in the control group would then be instructed not to offer treatment, while their colleagues in the treatment group would be allowed to assign treatment as they would do in real life, i.e. by taking private information on treatment heterogeneity into account. If outcomes depend (implausibly) only on P this will produce an unbiased estimate but it has lower statistical power than the regression approach. (The reason is that the RCT approach compares average outcomes at the level of program officers while the regression approach makes the comparison at the level of beneficiaries.)

In the general case, when outcomes depend on both P and X this approach is likely to fail. Program officers will have been assigned to different locations for particular reasons so that P and X are not independent at this level. The treatment and control groups would therefore differ in terms of X (in expectation) so that internal validity is destroyed. RCTs are then useless.

What does this mean in practice? First, the evaluator should decide whether the effect of the treatment is likely to differ between various beneficiaries. Next, he should carefully study the nature of the NGO intervention to determine whether the decision to assign "treatment" is influenced by private information about treatment heterogeneity. If, the answer to either question is negative then a standard RCT evaluation (with randomization at the level of beneficiaries) is appropriate. This is illustrated in Figure 2.

However, if

– there is no strong hierarchical control in the NGO: program officers have discretion to determine whether someone receives treatment and how much,
– and they base that decision in part on the differences they perceive between beneficiaries in terms of the effect of the treatment

then RCTs are not appropriate. In that case the evaluation should use observational data and estimate the TPE as indicated above.

Randomization is crucial in either case. In RCTs the assignment to treatment and control groups should of course be random. When observational data are used these should be from a random sample of the target population to ensure that the β-coefficients are appropriately weighted.

The problem of imperfect assignment is, of course, not specific to NGOs. But it is particularly likely to arise in NGO evaluations since NGOs typically work in a very flexible, highly decentralised way, leaving room for experimentation and improvisation at the grassroots level. That this has important implications for the way NGOs are evaluated seems to be insufficiently appreciated.

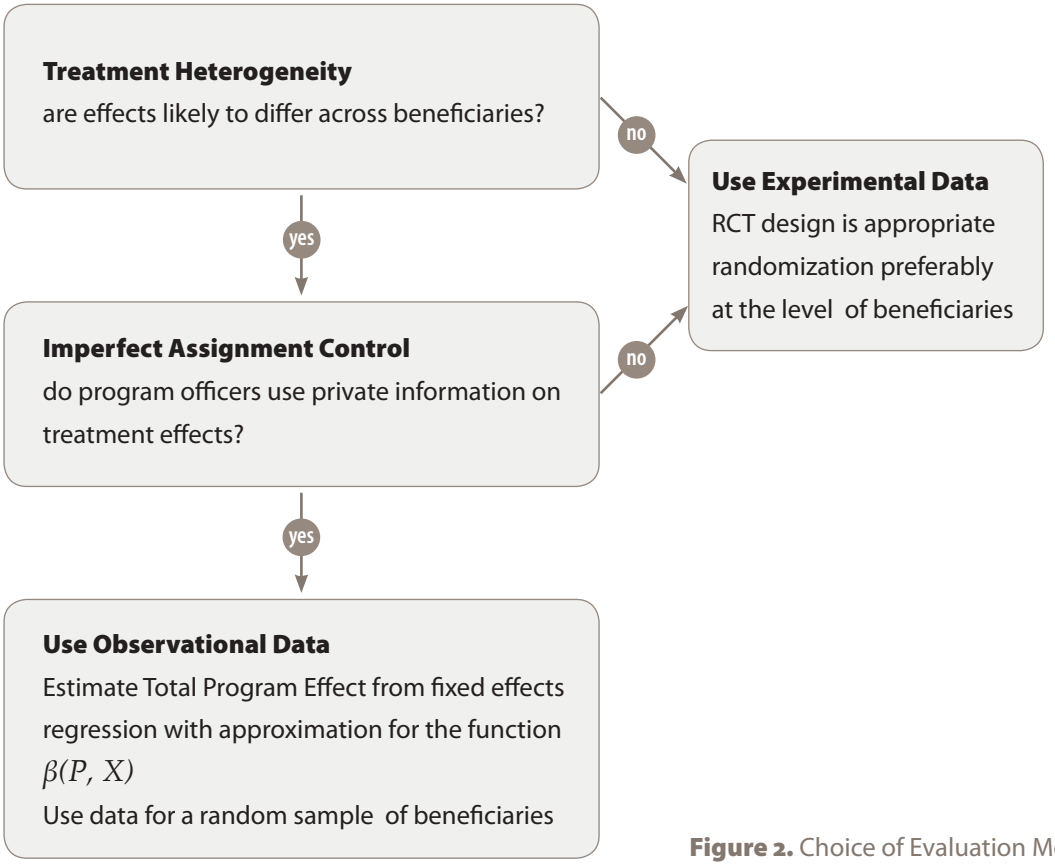

Figure 2. Choice of Evaluation Method

5. Conclusion

Modern impact evaluation methods such as RCTS and regression discontinuity designs have become extremely popular in development economics. There is some resistance to these methods in development NGOs. We have suggested that the claim that standard impact evaluation cannot be applied to NGOs is not without foundation.

We have discussed two cases where RCTs are indeed inappropriate. The first arises when an NGO aims to achieve its objectives indirectly, through advocacy which should result in policy changes. Depending on the coverage of the targeted policies this may leave little scope for an RCT evaluation. We have argued that nevertheless there remains considerable room for a rigorous evaluation design.

The second case arises when the organization's control is imperfect in the sense that decisions on the targeting of the program (actual treatment assignment or intention to treat) are taken by local staff (program officers) who can use private information on the effectiveness of the intervention for particular beneficiary households or communities. We have shown that this case (which is quite plausible in an NGO context) presents a very serious challenge for an evaluation. An RCT is unlikely to be appropriate in this context. We have shown how the problem can be surmounted in a fixed effect regression framework, using observational rather than experimental data. This approach explicitly deals with unobserved treatment heterogeneity.

References

- **AIID** (2011), 'Proposal for an Evaluation of Twaweza Activities', Amsterdam Institute for International Development", November 16.

- **Deaton, Angus** (2010), 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature*, vol. 28, pp. 424-455.

- **Elbers, Chris** and **Jan Willem Gunning** (2012), 'Evaluation of Development Programs: Using Regressions to Assess the Impact of Complex Interventions', Tinbergen Institute Discussion Paper 12-069/2.

- **Miguel, Edward** and **Michael Kremer** (2004), `Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities', *Econometrica*, vol. 72, pp. 159-217.

FERDI

Créée en 2003, la **Fondation pour les études et recherches sur le développement international** vise à favoriser la compréhension du développement économique international et des politiques qui l'influencent.

Contact
www.ferdi.fr
contact@ferdi.fr
+33 (0)4 73 17 75 30