UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2014/026

Evaluation of non-governmental development organizations

Chris Elbers and Jan Willem Gunning*

January 2014

Abstract: Randomized controlled trials (RCTs) are now widely used in development economics. However, their use is often resisted by non-governmental development organizations. The objections they raise differ between the three types of activities of such non-governmental organizations (NGOs): capacity building, advocacy, and service delivery. This paper discusses the objections and alternatives to RCTs for each type. RCTs might not be appropriate even for service delivery, the activity which would appear to be best suited to their use. This is because typically local NGO staff can use their discretion in selecting communities or individuals for participation in a service-delivery programme. A standard RCT does not mimic the use of private knowledge of local circumstances and can therefore be misleading.

Keywords: impact evaluation, randomized controlled trials, non-governmental development organizations
JEL classification: C21, C33, O22

* Both authors VU University Amsterdam, Tinbergen Institute; c.t.m.elbers@vu.nl ; j.w.gunning@vu.nl

This paper was prepared for the‘Oxford Handbook of Africa and Economics’ authors’ conference in Beijing, 8-10 December 2013, implemented with UNU-WIDER.

ISSN 1798-7237 ISBN 978-92-9230-747-9 https://doi.org/10.35188/UNU-WIDER/2014/747-9

Typescript prepared by Anna-Mari Vesterinen at UNU-WIDER.

UNU-WIDER gratefully acknowledges the financial contributions to the research programme from the governments of Denmark, Finland, Sweden, and the United Kingdom.

The World Institute for Development Economics Research (WIDER) was established by the United Nations University (UNU) as its first research and training centre and started work in Helsinki, Finland in 1985. The Institute undertakes applied research and policy analysis on structural changes affecting the developing and transitional economies, provides a forum for the advocacy of policies leading to robust, equitable and environmentally sustainable growth, and promotes capacity strengthening and training in the field of economic and social policy-making. Work is carried out by staff researchers and visiting scholars in Helsinki and through networks of collaborating scholars and institutions around the world.

UNU-WIDER, Katajanokanlaituri 6 B, 00160 Helsinki, Finland, wider.unu.edu

The views expressed in this publication are those of the author(s). Publication does not imply endorsement by the Institute or the United Nations University, nor by the programme/project sponsors, of any of the views expressed.

1 Introduction

How do we know that a non-governmental development organization is effective? Evaluation reports written by commercial consultants typically describe what a non-governmental organization (NGO) has done and how outcomes have changed over time, *implying* that these changes can be attributed to the NGO's actions. This has never been convincing, but only recently has the disenchantment with such before-after comparisons in most evaluation reports become widespread.[1]

There is now a surge of interest in using rigorous impact evaluation methods, such as randomized control trials (RCTs), to assess what works in development.[2] These experimental methods are now commonly applied to evaluate NGOs in Africa. Indeed, one of the most famous papers in the RCT literature, the 'deworming paper' of Miguel and Kremer (2004), describes an RCT evaluation of an NGO in Kenya.

While these methods are now widely used, they are sometimes resisted, particularly by development NGOs. Often this reflects no more than an irrational aversion to rigour and quantitative methods. But, NGOs can also object on the grounds that some NGO activities do not lend themselves to impact evaluation methodologies, such as RCTs. This position deserves to be taken seriously. It is the main focus of this paper.

NGOs active in development are engaged in capacity building, advocacy, service delivery, or some combination. The objections raised to the use of RCTs differ between these three types of activities. In the case of capacity building, e.g., a training programme for the staff of a partner organization, the objection is that the treatment group is typically very small and in any case not randomly selected. In the case of advocacy, the objection is that there is obviously little scope for RCTs when other NGOs target the same policies or when the policies the NGO is trying to get changed cover large geographical areas. In the extreme case of *national* policies, there clearly is no scope at all for RCTs—since everybody is affected, the impact of the policy cannot be identified. The third type of activity, service delivery, would seem to be best suited to RCTs. However, we will suggest that even in this context RCTs may not be appropriate. The reason is that in many NGOs, local staff can use their discretion in selecting communities or individuals for participation in, say, a sanitation programme. A standard RCT will then be misleading, since it cannot mimic the use of private knowledge of local circumstances in such targeting.

We consider these three cases in turn in Sections 2, 3, and 4. Section 5 concludes the paper.

2 Capacity building

Many international NGOs are engaged in capacity building for partner organizations, e.g., African NGOs. Such activities are usually evaluated with qualitative methods.[3] Often two

[1] This paper draws on Elbers and Gunning (forthcoming); and Gunning (2012, 2012a).

[2] Duflo et al. (2008); as well as Banerjee and Duflo (2011) describe such methods; Ravallion (2009, 2012); and Deaton (2010) are well known critiques of this approach.

[3] The Netherlands' government supports Dutch development NGOs through a co-financing programme. The current phase of that programme (called MFS II) is currently being assessed in a massive evaluation of a large sample of projects, stratified by country, NGO, and type of activity. The sample covers six countries, including four

reasons are invoked for not using formal methods (e.g. Barrett et al. 2012). First, capacity building usually involves only a very small number of partner organizations (perhaps only a single one) so that the treatment group is too small for statistical analysis. Second, many NGOs are *sui generis*, so no meaningful control group can be identified. (There are dozens of NGOs in many African countries focusing on issues such as maternal health, enrolment of girls, or microfinance, and hence, this objection seems to be exaggerated).

Nevertheless, there is some scope for going beyond informal, qualitative methods, provided one is willing to judge the capacity that is being built in terms of *actual* rather than *potential* change in the way a partner organization functions. In other words, instead of assessing informally what an organization was previously unable to do but now can do, one assesses whether the organization actually functions better than a similar organization which was not subjected to capacity building. That assessment is based on how the intended beneficiaries are affected by the NGO. Hence, one would compare the beneficiaries (say a group of women targeted by a women's empowerment programme) of one partner organization with the beneficiaries of other NGOs (in this case, also aiming at women's empowerment).

This does not, of course, address the second objection, but it does address the first one. By focusing not on organizations but on their beneficiaries, one may well be able to reach a reasonable sample size and in fact follow a diff-in-diff approach.

3 Advocacy: achieving change indirectly

As already noted in the Introduction, evaluating the advocacy activities of an NGO is inherently difficult for two reasons. First, usually there are other institutions, including other NGOs, aiming at a same policy change. This makes attribution difficult: If the intended policy change occurs, this might not be the result of one NGO's efforts. Second, when the policy in question is a national one, a convincing counterfactual cannot be constructed. In particular, one cannot assess the effect of the policy change by comparing randomly selected locations with and without the policy since *all* locations are affected, although not necessarily to the same extent. Hence either the evaluator cannot say how much difference the policy change made or he cannot attribute that change to the NGO in question. Perhaps all he can say is that if the policy was *not* changed there is a *prima facie* case for concluding that the NGO was not effective.

Often, however, one can go further. We illustrate this with an example of an East African NGO, Twaweza.[4] Twaweza, based in Dar es Salaam, is active in three countries: Tanzania, Kenya, and Uganda. The organization is involved in numerous activities. It is particularly active in national debates on the quality of public services and it lobbies for policy changes in the health and education sectors. Clearly, the two limitations on the evaluation of advocacy apply to this part of Twaweza's work.

in Africa: the Democratic Republic of Congo (DRC), Ethiopia, Liberia, and Uganda. Barrett et al. (2012) describe the evaluation methodology used by most country teams for the capacity-building projects. Essentially this involves an informal assessment of success along various dimensions where the assessment is done by the staff of the NGO involved, external stakeholders, and the evaluators themselves. There is no comparison with other organizations (e.g., a control group of comparable NGOs, which were not involved in the capacity building by the Dutch NGOs). In effect, this shifts the attribution problem to the respondents: They have to decide informally to what extent changes over time can be attributed to the capacity building. Such evaluation approaches are quite common.

[4] See http://www.twaweza.org.

But another part of Twaweza's work is aimed at *indirect* policy change. The idea is to change, through advocacy, the information available at the local level, to trigger local collective action through that information, and thereby to improve development outcomes. This can be seen as a theory of change with three steps:

(1) Advocacy changes the information available at the local level. For example, people in rural areas learn through a radio programme that children learn very little in schools.

(2) This information leads to collective action. For example, parents complain to the school principal or to their MP about the quality of education.

(3) This action leads to better outcomes. For example, the principal reacts by reducing teacher absenteeism, which results in better learning.

Note that this implies three different tests, and that the theory must be rejected if it fails any of them. In that sense, there is no need to test steps 2 and 3 if the evaluator has already found that the theory does not pass the first test. However, evaluations are increasingly seen as inputs into a learning process rather than as one-off assessments. A learning organization would definitely want the evaluator in this example to proceed with testing steps 2 and 3 after a negative finding in step 1. For if the theory passes those tests, then the NGO would have learned that it needs to change the way it gets information to villages and that it is worth doing so since if information reaches the villages it lead, as intended, to collective action and better outcomes.

Any test of the first step is clearly asymmetric: If no such information reaches the village, then the theory obviously fails the test, but if it does arrive, then one cannot confidently attribute this to the NGO, notably because many other organizations may be active in a similar way.

How can one test the first step? RCTs are of no use here, even if the NGO uses locationally specific channels, e.g. radio programmes which can be heard only in certain parts of the country. The reason is that such a geographic limitation does not involve randomization so that it cannot be used as the basis for defining treatment and control groups. In general villages, where the programme can be heard will differ in many other respects from other villages.

The alternative is to use a team of respondents in a representative sample of villages and to ask them frequently what people talk about, in particular, if it is related to health and education, and what the source of their information is.[5] An ongoing (2012–14) evaluation of Twaweza, in which the authors of this paper are involved together with colleagues at the Institute of Rural Development planning (IRDP) at the University of Dodoma, uses this approach. Respondents in a sample of 250 locations throughout Tanzania are interviewed by mobile phone at high frequency (every three weeks) over the two-year period. The interviewers, who are IRDP staff members, guide the respondents through a semi-structured interview. Respondents are asked whether there is any news (since the previous conversation) regarding health and education. If this is the case, the interviewer asks a series of questions to establish the source of that information: a visitor from another village; a politician who gave a speech; a religious leader who came to the village; a radio or TVprogramme; and so on. At a later stage, the interviewer asks more leading questions, e.g. whether the respondent knows of a particular radio programme (e.g. one sponsored by Twaweza, although the respondent need not know that).

[5] This can be seen as a mix of the traditional anthropologist's approach (who finds out what people talk about in the village he lives in) and the economist's (who wants to generalize and therefore collects data in a representative sample of villages). The anthropologist's information is typically richer (but limited to quite a small sample), and the economist's superficial (but suitable for generalization).

This procedure can generate three types of evidence:

(a) Information which Twaweza promotes does not reach the village or only does so rarely;

(b) The information does reach the village, but this cannot be attributed to Twaweza's efforts, e.g. because there are other organizations who also provide information about poor quality of teaching in primary schools;

(c) The information does reach the village and it is clear that this can be attributed to Twaweza, e.g. because the respondent identifies a Twaweza-supported radio programme as their source.[6]

The phone panel also collects data on collective action, e.g. parents challenging teachers, holding politicians to account, or engaging in self-help schemes. In this case too, the respondent is first given an opportunity to give the informationvoluntarily without any prompting from the interviewer. If they do not do so, the interviewer follows-up with more 'leading' questions in order to establish whether actions not yet mentioned by the respondent have in fact taken place.

Now consider the second step in the theory of change. Here the question is whether information about, say, health and education (irrespective of whether it comes from Twaweza) leads to collective action. There is no particular reason why this should be so. The theory of change implies that there is only one reason why people do not act collectively: They do not know how bad the situation is. In that case, the provision of credible information about that situation will indeed trigger collective action. But, there are many reasons why this need not be the case. The information may contain no news for the parents: They may already know very well how badly their children are taught but have decided they cannot do anything about it. Collective action is notoriously difficult to organize since there are strong incentives for free riding behaviour, people may be afraid of retaliation by people in power, they might be convinced that action cannot succeed and so on.[7]

There is some scope for using RCTs to test whether information leads to collective action and in the case of Twaweza there have been several such studies. Closest to the spirit of Twaweza's theory of change is the study by Lieberman et al. (2013). The study used an RCT to investigate the effect of providing information to 20 households in 30 randomly selected villages in Kenya. Children in the treatment households were tested and parents were informed about the literacy and numeracy skills of each child. They were also provided with a list of actions they could take to improve their child's skills. This included helping at school, attending a parents-teacher meeting, or discussing their child's performance with the teacher. The RCT evidence in this study is very clear: One cannot reject the hypothesis that information triggers neither such public action, nor any private action (such as helping children with their homework). Essentially the provision of information has *no* effect.

The alternative to such RCT is to use observational data. In the case of the Twaweza, evaluation the high frequency cell phone interviews provide data on both information received and any collective action undertook. These data can be used in a regression analysis to assess whether information (of all types, i.e. not necessarily information provided by Twaweza) leads to action.

[6] At the time of writing, much of the evidence is of the first type, but this is a very preliminary finding.

[7] See Lieberman et al. (2013), Figure 1.

The obvious objection is that in such a regression 'information' is endogenous: There may be something special about villages that pick up (or report) information and that 'something' might also make them more likely to engage in action. However, such omitted variables are likely to be time-invariant so that a diff-in-diff approach (which is feasible since the data are panel data) can deal with the resulting omitted variable bias.

Clearly, the communication strategy adopted by an NGO, such as Twaweza, may imply that information is more likely to arrive in those villages where it can trigger effective collective action. This non-random allocation is part of the programme's effectiveness and it would therefore be wrong to eliminate it, as an RCT would do.[8] However, this case is also problematic for a standard diff-in-diff regression approach: it is plausible that there are differences between locations in the effect of information on collective action (treatment heterogeneity) and that these differences are correlated with the assignment (the information picked up in the village). This introduces a type of endogeneity that cannot be dealt with by double differencing. However, the method of Elbers and Gunning (forthcoming) can be used in this case to obtain a consistent estimate of the impact of the programme, taking into account this correlation.[9]

For the evaluation of the second step in the theory of change, one can choose between two quite different approaches, relying on experimental data in the one case and on observational data in the other. The RCT approach has the advantage that it is far easier to implement: Setting up a large phone panel of rural respondents and maintaining it for a long period is quite difficult. The RCT approach also has a disadvantage as it focuses on a very specific form of information provision, whereas an NGO may use various channels (as does Twaweza) and change their use over time. A narrowly designed RCT might therefore raise external validity concerns which need not arise in the regression approach.

The effect of collective action, the third step in the theory of change, can also be addressed with regression analysis. In the Twaweza evaluation a baseline household survey was conducted in 2012 in the same 250 villages which are used as the sample for the phone panel. This survey collected data on household and community characteristics and also on outcomes, notably quality indicators for health and education services as well as on collective action. The survey will be repeated, probably in 2014. Changes in outcomes (the differences between the endline and baseline surveys) can then be regressed on various explanatory variables from the phone panel, including measures of public action in the intervening years. As before, omitted variables will introduce endogeneity which can be eliminated by differencing. In this case too, treatment effects are likely to be correlated with the extent of collective action. Again, this type of endogeneity can be dealt with by using the approach of Elbers and Gunning (forthcoming), which will be explained at length in the next section.

It is often suggested that advocacy activities of NGOs do not lend themselves to rigorous evaluation. This position is defensible if those activities are to be evaluated as an integrated whole so that the entire chain, from the advocacy itself, through responses from citizens to the resulting changes in outcomes, must be evaluated. In this case, finding a credible counterfactual will rarely be feasible. However, 'perfect' should not be the enemy of 'good': A substantial part of the theory of change *does* lend itself to rigorous analysis in the example. This involves 'opening

[8] We do not wish to suggest that this invalidates the Lieberman et al. (2013) study. Whether treatment heterogeneity (and the correlation with assignment) is important in the case of Twaweza remains to be seen. This question can be answered once the regression analysis can be performed. Clearly, the two approaches may reach the same conclusion that information does not induce collective action.

[9] The method is explained in Section 4.

the black box' by testing various components of the theory of change separately. Both the effect of information on collective action and the effect of the collective action on development outcomes can be investigated rigorously. We have indicated that in some cases RCTs are not suitable for this purpose and that instead regression methods should be applied to observational data (as opposed to the experimental data generated by an RCT).

We expect that the type of endogeneity we have focused on (arising from correlation between programme assignment and heterogeneous treatment effects as a result of the discretionary powers of the programme officer) will come to be seen as a central problem in evaluations, certainly those of NGOs. In this situation, RCTs are less suitable than is commonly assumed but a regression-based alternative approach exists. The issue is not specific to advocacy. We will see in the next section that it also very much affects the evaluation of NGOs involved in service delivery.

4 Imperfect control in service delivery

Of the three types of NGO activities we have distinguished, service delivery lends itself best to an RCT evaluation. However, an ex ante evaluation is sometimes not feasible: Donors often set up an evaluation when project activities have already started so there is no longer scope for randomization. Since the treatment group has already been selected, the evaluation is of the ex post type. In this case, the evaluator can construct a control group using matching techniques to ensure its comparability with the given group of beneficiaries. Instead of RCTs, the evaluation then applies diff-in-diff techniques to baseline and endline data for the two groups of beneficiaries.[10] This is a sensible approach but may produce biased estimates when the NGO's control over programme participation is imperfect. In this section we focus in that case.

In many NGOs, policy makers at headquarters (HQ) set policies only in quite general terms. They may, for example, decide that a particular educational programme is to be implemented in villages which are poor according to some criterion. Specific decisions on programme participation are often left to staff in the field (programme officers) who can exercise considerable discretion within the wide guidelines of the HQ policy. In particular, programme officers can use private information on where the programme is likely to be most effective.

Such delegation of decision making power is, of course, entirely sensible. However, it presents a problem for an evaluation since it creates a correlation between treatment effects and assignment which implies endogeneity.[11] The problem arises if both of the two conditions are satisfied: The effect of the programme differs between individuals or locations (treatment heterogeneity) and the programme officer bases his decisions on what communities or which individuals are able to participate in the programme at least in part on those differences. Since this conjunction is quite common in NGOs the issue needs to be taken seriously.

[10] In the evaluation of the Netherlands government's support for Dutch development NGOs (cf. Footnote 3), this approach has been adopted by all country teams to assess the impact of the programme in terms of the Millennium Development Goals (MDGs).

[11] This is an example of what Heckman et al. (2006) termed essential heterogeneity. See also Ravallion (2011). Note that the correlation can arise in two ways, depending on whether the behavioural response to the treatment heterogeneity comes from the programme officer or the beneficiary. We focus on the former case, Ravallion (2011) on the latter, i.e., the case of selective take up.

To see the problem, it is helpful to write the evaluation in regression form. Denote the outcome variable (e.g. a measure of learning in an education intervention) as *y*, a vector of intervention variables (e.g. the number of school books and the number of trained teachers in a particular school) as *P*, a vector of control variables as *X* and the error term as ε. We eliminate the effect of (time-invariant) unobservables by taking differences.[12] This gives the regression equation as:

$$\Delta y_i = \alpha\Delta X_i + \beta_i\Delta P_i + \gamma + \Delta\varepsilon_i = \alpha\Delta X_i + \overline{\beta}\Delta P_i + \gamma + [(\beta_i - \overline{\beta})\Delta P_i + \Delta\varepsilon_i] \qquad (1)$$

where *i* denotes the unit of observation (e.g., a school or a child) and α, β_i, γ are parameters.[13] The effect of the treatment is measured by β and the notation β_i indicates that we allow for treatment heterogeneity: the treatment effect can differ across *i*.

We assume that in (1) the controls *X* are exogenous. There is, however, another source of endogeneity: Who can participate in the NGO programme (or is offered the opportunity to participate) is decided by local staff—the programme officers base their decisions on their estimates of β_i, the effectiveness of the programme at the individual level. This private information need not be correct: We only assume that the programme officer's estimates, and hence, decisions are correlated with the individual treatment effects.

Since β_i and ΔP_i are correlated, it is clear from the second part of equation (1) that ordinary least squares (OLS) estimation would be inappropriate: ΔP_i is correlated with the error term in square brackets. In the terminology of Heckman et al. (2006) this is the case of essential heterogeneity which we already encountered in the previous section. The usual remedy of instrumental variable estimation now *necessarily* fails: As can be seen from Equation (1), any variable which is correlated with ΔP_i must also be correlated with the error term in square brackets. Therefore a variable which satisfies both requirements for a valid instrument does not exist. Hence, an evaluation using standard regression methods on observational data runs into an insurmountable endogeneity problem.

In general, an RCT will fare no better. In non-technical terms: An RCT would be invalid in this context since it would choose participants (e.g. schools in the treatment group) randomly whereas in actual practice they would be chosen non-randomly by programme officers. The RCT is misleading since it does not mimic how 'treatment' (or at least the offer of treatment) is assigned in practice. It could produce a 'false negative', concluding that the intervention is not effective when in fact it is.

In this situation an RCT misses a key characteristic of the NGO: Its reliance on the private information of programme officers. This is a situation where standard evaluation methods are indeed inappropriate for NGOs.[14]

[12] This implies the assumption of parallel trends.

[13] Note that γ picks up the effect of a linear time trend.

[14] We are not suggesting that this problem is specific to NGO evaluations, only that it is quite likely to occur in that context.

The solution might seem to apply an RCT at a higher level, by randomizing at the level of programme officers rather than beneficiaries (schools in this example). A drawback is that this method involves loss of statistical power.[15] The method is inefficient since the programme will, obviously, not be offered to any of the schools in the control group. Additionally, since programme officers can exercise their discretion, it will also not be offered to some members of the treatment group. As a result, the proportion of actually treated cases in the trial will be too small.

A much more serious problem is that randomization over programme officers may destroy internal validity of the RCT. This would arise if programme officers were stationed in particular areas of the country on the basis of, say, their membership of an ethnic group or their knowledge of relevant farming methods. This would lead to a correlation between characteristics of the programme officer and *X* and thereby between *P* and *X* at an individual level. In this case, randomization in a small sample of programme officers would result in systematic differences between the treatment and control groups: It is unlikely that randomization over *P* will also achieve randomization over *X*. Hence, internal validity is no longer guaranteed: Differences between the groups may reflect the differences in *X* rather than the treatment.

If neither RCTs nor standard regressions can be used, then evaluation would appear to be impossible. In Elbers and Gunning (forthcoming), we propose a solution to this fundamental problem. This involves collecting observational data[16] on the variables in Equation (1), e.g., in the case of a sanitation programme, changes in outcomes,such as health and participation, and in other determinants of the outcomes, such as household income. These data can be used in a regression to estimate the effect of the programme in a way which avoids the endogeneity problem. We are interested in an estimate of the total programme effect

$$E\beta_i \Delta P_i,$$

the expected value (in the population) of the intervention and its individual effect, *taking into account the correlation between the two*.[17]

The trick is to allow explicitly for a dependence of the assignment on the individual treatment effects (which we cannot observe) and the control variables (which are observed):

$$\Delta P_i = f(\beta_i, \Delta X_i).$$

If this function can be inverted, we have:

$$\beta_i = g(\Delta P_i, \Delta X_i). \qquad (2)$$

We can therefore rewrite the regression equation as

$$\Delta y_i = \alpha \Delta X_i + E(\beta_i \mid \Delta X_i, \Delta P_i)\Delta P_i + \gamma + \omega_i, \qquad (3)$$

[15] This is analysed in detail in the supplemental material of Elbers and Gunning (2014), but the key point is straightforward.

[16] That is: administrative or survey data rather than experimental data.

[17] Note that we cannot estimate a selection equation: we observe ΔP_i but not β_i.

where $\omega_i = \Delta\varepsilon_i + (\beta_i - E(\beta_i \mid \Delta X_i, \Delta P_i))\Delta P_i$ and this is uncorrelated with ΔX_i and ΔP_i.

Using a polynomial approximation we can write the term $E(\beta_i \mid \Delta X_i, \Delta P_i)$ as a series of terms in ΔX_i and ΔP_i. For example, for a linear approximation:

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

Substituting this in (3) gives

$$\Delta y_i = \gamma + \theta_1 \Delta X_i + \theta_2 \Delta P_i + \theta_3 \Delta X_i \otimes \Delta P_i + \theta_4 \Delta P_i \otimes \Delta P_i + \omega_i \tag{4}$$

Using observational data, this equation can be estimated using OLS.

The effect of the programme, which we call the total programme effect (TPE), can now be estimated by multiplying the regression coefficients of all terms involving ΔP with the sample means of the regressors:

$$T\hat{P}E = \hat{\theta}_2 \overline{\Delta P_i} + \hat{\theta}_3 \overline{\Delta X_i \otimes \Delta P_i} + \hat{\theta}_4 \overline{\Delta P_i \otimes \Delta P_i} \tag{5}$$

where bars denote the sample means.[18] This gives a consistent estimate of the total programme effect provided the observational data are from a representative sample (or can be reweighted).

In practice, this means that one regresses Δy_i on ΔX_i, ΔP_i and their interactions with ΔP_i and collects all terms involving ΔP_i to calculate the total programme effect. Since the estimated TPE is linear in the $\hat{\theta}$ parameters, its standard error can be obtained straightforwardly from the covariance matrix of the OLS-coefficients.

It is useful to stress why this approach provides a solution. The expected value of the treatment effect $E\beta_i$ is a parameter that can be obtained in an RCT (with randomization over beneficiaries) but not in a regression (because of essential heterogeneity). However, irrespective of whether it can be estimated or not, it is of no use in an evaluation since in practice programme assignment is *not* random.

The total programme effect is the expected value of the product $\beta_i \Delta P_i$. Clearly, this is equal to the product of the expected value of the two components only if these are independent, but in this case they are correlated (because the assignment, decisions have been delegated to the programme officers). Our approach cuts this Gordian knot by focusing directly on the product, the TPE itself, rather than on its components.

[18] Obviously, to identify θ_4 a restriction on parameters like $\theta_{4,k\ell} = \theta_{4,\ell k}$ is required.

It is instructive to consider the special case where $D_i = \Delta P_i$ is a binary variable taking the value 1 for the treatment group and 0 for the control group. Equation (4) now reduces to

$$\Delta y_i = \gamma + \theta_1 \Delta X_i + \theta_2 D_i + \theta_3 D_i \Delta X_i + \omega_i$$

since in the binary case $D_i^2 = D_i$. The total programme effect will then be estimated as

$$\hat{\text{TPE}} = \hat{\theta}_2 \overline{D_i} + \hat{\theta}_3 \overline{D_i \Delta X_i} \,. \tag{6}$$

As before, this shows that when the sample is representative, the sample means can be used to construct the total programme effect. The interaction term in (6) avoids the bias resulting from correlations between treatment effects and either programme participation or controls.

Standard diff-in-diff regressions usually omit this interaction term (e.g. Khandker et al. 2009; Almeida and Galasso 2010). In addition, our method can also be used in the general case when the intervention is multi-dimensional and at least some of its components are multi-valued rather than binary.

How can this method be applied in practice? The evaluator should first ascertain how the NGO works. If programme officers have discretion to determine whether someone receives treatment (and how much) and they base that decision in part on the differences they see between beneficiaries in the effect of the treatment then RCTs are inappropriate. The evaluation should then use observational data and employ the method we have described to estimate the effectiveness of the NGO.[19]

NGOs, in Africa and elsewhere, have often reacted in extreme ways to the recent enthusiasm for RCTs. Some have embraced the methodology uncritically in the hope that the ‘rigour’ of RCT evaluations will give their work greater credibility. Other NGOs have rejected the methodology without fully understanding it, often arguing that quantitative methods miss the essence of their work. Neither reaction is appropriate. Much of the work of NGOs involved in service delivery (e.g. programmes in health and education) is likely to be of the type which we have characterized as that of programme officer discretion. In that case, RCTs indeed miss the essence of the NGOs work and their results can be misleading.[20] Fortunately, the problem can be addressed with the method we have proposed. As data availability continues to improve, this will become easier to apply.

We do not wish to suggest that there is no role for RCTs in this case. However, their role is to produce a convincing ‘proof of principle’. Beyond that, they reach a natural limit: If treatment heterogeneity is important, then an RCT cannot predict the effect of a programme which in practice will be assigned in a way different from the assignment in the RCT.

[19] One can test whether those variables, which would normally not be included in the regression (the interactions of treatment variables with themselves and with the controls), are jointly significant. When this test indicates that treatment heterogeneity cannot be rejected, it is advisable to calculate the TPE. This simple test should in our view always be performed. Allowing for treatment heterogeneity may turn out to be an unnecessary precaution but this is small price to pay for avoiding a biased estimate of the effect of the NGO’s activities.

[20] Clearly, the problem is not the quantitative nature of an RCT. Indeed, any qualitative method faces the same issues.

5 Conclusion

NGOs are a major presence in African economies. In many ways, they attempt to do what governments are not able or willing to do. Their effectiveness has only recently begun to be assessed rigorously, e.g. by using RCTs to evaluate their activities.

In this paper, we have argued that when deciding what can and cannot be evaluated using RCTs, it is important to distinguish between three types of activities of development NGOs: capacity building, advocacy, and service delivery. (Of course, many NGOs are involved in some combination of the three types.) It is generally accepted that service delivery activities can be evaluated rigorously. Nevertheless, it is commonly suggested that there is very little scope for rigorous evaluation of capacity-building and advocacy activities of NGOs. We have suggested that this position is exaggerated.

We have also argued that the case for RCT evaluations is overstated. The very nature of NGOs with their non-hierarchical organization, leaving considerable scope for discretion to staff in the field, makes it unlikely that RCTs will produce useful estimates, at least in situations where treatment heterogeneity is important. While the issue of essential heterogeneity may sound as an esoteric concern of econometricians, the problem is likely to be pervasive, certainly in the case of NGOs. We therefore very much agree with Ravallion's statement that: 'Essential heterogeneity is such an intuitively plausible idea that the onus on analysts should be to establish *a priori* grounds why it does not exist'.[21] This suggests greater reliance on observational instead of experimental data. (As data availability improves, this will become easier to implement.) We have indicated how observational data can be used to evaluate NGO programmes when standard RCT methods would fail because of essential heterogeneity.

References

Almeida, R.K. and E. Galasso (2010). 'Jump-starting Self-employment? Evidence for Welfare Participants in Argentina'. *World Development*, (38): 742–55.

Banerjee, A. and E. Duflo (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: Public Affairs.

Barrett, J., P. Bilinsky, C. Desalos, D. Hilhorst, E. Kamphuis, C. Kusters, D. Klaver, and A. Groot Kormelinck (2012). 'Evaluation Methodology MFS II Country Evaluations; Capacity of Southern Partner Organisations (5 Cs) Component; Evaluation Methodology for DRC, Ethiopia, India, Indonesia, Liberia and Uganda'. Wageningen: Center for Development Innovation.

Deaton, A. (2010). 'Instruments, Randomization, and Learning about Development'. *Journal of Economic Literature*, (48): 424–55.

Duflo, E., R. Glennerster, and M. Kremer (2008). 'Using Randomization in Development Economics Research: A Toolkit'. In T.P. Schultz and J. Strauss (eds), *Handbook of Development Economics*. Amsterdam: North-Holland.

Elbers, C. and J.W. Gunning (forthcoming). 'Evaluation of Development Programs: Randomized Controlled Trials or Regressions?'. *World Bank Economic Review*.

[21] Ravallion (2011: 9, italics and underlining in the original). Recall, however, that Ravallion is concerned about self selection into treatment rather than the case we have focused on where assignment is decided by the programme officer who uses private information on treatment effects.

Gunning, J.W. (2012). 'How Can Development NGOs be Evaluated?'. FERDI Working Paper 51. Clermont-Ferrand: FERDI.

Gunning, J.W. (2012a). 'Evaluating Development NGOs'. Policy Brief 56. Clermont-Ferrand: FERDI.

Heckman, J.J., S. Urzua, and E. Vytlacil (2006). 'Understanding Instrumental Variables in Models with Essential Heterogeneity'. *Review of Economics and Statistics*, (88): 389–432.

Khandker, S.R., Z. Bakht, and G.B. Koolwal (2009). 'The Poverty Impact of Rural Roads: Evidence from Bangladesh'. *Economic Development and Cultural Change*, (57): 685–722.

Lieberman, E.S., D.N. Posner, and L.L. Tsai (2013). 'Does Information Lead to More Active Citizenship? Evidence from an Education Intervention in Rural Kenya'. MIT Political Science Department Research Paper No. 2013–2. Cambridge, MA: MIT.

Miguel, E. and M. Kremer (2004). 'Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities'. *Econometrica*, (72): 159–217.

Ravallion, M. (2012). 'Fighting Poverty One Experiment at a Time: A Review of Abhijit Banerjee and Esther Duflo's Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty'. *Journal of Economic Literature*, (50): 103–14.

Ravallion, M. (2011). 'On the Implications of Essential Heterogeneity for Estimating Causal Impacts Using Social Experiments'. World Bank Staff Working Paper 5804. Washington, DC: World Bank.

Ravallion, M. (2009). 'Evaluation in the Practice of Development'. *World Bank Research Observer*, (24): 29–53.