TI 2012-081/2
Tinbergen Institute Discussion Paper

Evaluation of Development Programs: Using Regressions to assess the Impact of Complex Interventions

Chris Elbers
Jan Willem Gunning

Faculty of Economics and Business Administration, VU University Amsterdam, AIID, and Tinbergen Institute.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Evaluation of Development Programs:

Using Regressions to Assess the Impact of Complex Interventions[1]

Chris Elbers and Jan Willem Gunning

VU University Amsterdam, Tinbergen Institute and AIID

Revised July 2012

Abstract

There is a growing interest in extending project evaluation methods to the evaluation of programs: complex interventions involving multiple activities. In general a program evaluation cannot be based on separate evaluations of its components since interactions between the activities are likely to be important. We propose a measure of program impact, the total program effect (TPE), which is an extension of the average treatment effect on the treated (ATET). Regression techniques can be applied to observational data from a representative sample to estimate the TPE for complex interventions in the presence of selection effects and treatment heterogeneity. As an example we present an estimate of the TPE for a rural water supply and sanitation program in Mozambique.

Estimating the TPE from randomized controlled trials would appear to be an alternative; however, the scope for using RCTs in this context is limited.

JEL Codes: C21, C33, O22

keywords: program evaluation; randomized controlled trials; policy evaluation; treatment heterogeneity; budget support; sector-wide programs; aid effectiveness

[1] We are grateful to Remco Oostendorp, Menno Pradhan, Martin Ravallion, Finn Tarp and seminar participants in Amsterdam, Duisburg-Essen, Namur, Oxford and Paris for comments on previous versions.

Evaluation of Development Programs:

Using Regressions to Assess the Impact of Complex Interventions

1. Introduction

Experimental techniques for impact evaluation presuppose that the intervention is well-defined: the "project" is limited in space and scope (e.g. Duflo *et al.*, 2008). However, governments, NGOs and donor agencies are often interested in evaluating the effect of a program consisting of heterogeneous interventions such as sector-wide health or education programs (De Kemp *et al.*, 2011). Program evaluation faces two complications. First, a dichotomous distinction between treatment and control groups is usually impossible. For example, a program in the education sector may involve activities such as school building, teacher training and supply of textbooks. Typically *all* communities are affected in some way by the program, but they may differ dramatically in what interventions they are exposed to and the extent of that exposure. Secondly, in a program the interventions are typically implemented at various administrative levels so that the policy maker has only imperfect control over actual treatment.

The impact of such a program cannot simply be calculated on the basis of the results of randomized controlled trials (RCTs). This would run into well known problems of external validity (Bracht and Glass, 1968, Rodrik, 2008, Ravallion, 2009, Banerjee and Duflo, 2009, Deaton, 2010, Imbens, 2010) even if the program involved only a single intervention. In addition, if the program involves multiple interventions and interactions are important then it is not even clear how to assess the overall impact of the program, even if individual components of the program have all been evaluated by means of RCTs. We will argue, however, that regression techniques can be used for evaluation in a sector-wide context. This involves drawing a

representative sample of beneficiaries (e.g. households, schools, communities) and collecting data on the combination of interventions experienced by each beneficiary together with other possible determinants of the outcome variables of interest. Regression techniques can then be used to estimate the impact of the various interventions.[2] In this paper we generalize this approach by allowing for treatment heterogeneity and propose an estimate of aggregate program impact.

Clearly, the intervention variables included in the regression as explanatory variables may be endogenous. For example, an unobserved variable such as the political preferences of the community may affect both the impact variable of interest and the intervention. In addition, the impact of the intervention may differ across beneficiaries and the allocation of interventions across beneficiaries may in part be based on such treatment heterogeneity, either through self-selection or through the allocation decisions of program officials. In either case the intervention variables would be endogenous. We will argue that to the extent that endogeneity is the result of treatment heterogeneity ("selection on the gain", Heckman, 1997, Heckman *et al.*, 2008) one should *not* correct for it since the resulting selection bias is part of the program impact.

The rest of the paper is organized as follows. In the next section we propose a measure of program impact, the total program effect (TPE), which extends the average treatment effect on the treated (ATET). We then consider two complications: correlation between program variables and the controls in section 3 and spillover effects in section 4. In section 5 we investigate whether estimating the TPE using RCTs is an alternative. The scope for RCTs is limited, notably when in the program assignment is imperfectly controlled and correlated with unobservables. We illustrate the approach in Section 6 by estimating the TPE for a program in Mozambique involving water supply and sanitation training interventions. Section 7 concludes.

[2] This approach is discussed in White (2006) and Elbers *et al.* (2009).

2. The Total Program Effect (TPE)

Consider the following model:

$$y_{it} = f(X_{it}) + g_i(P_{it}) + \eta_i + \varepsilon_{it} \qquad (1)$$

where y measures an outcome of interest, in this paper taken to be a scalar; $t = 0, 1$ is the time of measurement; and $i = 1,...,n$ denotes the unit of observation (*e.g.* households or locations). P denotes a vector of the interventions to be evaluated and X a vector of observed controls.[3] The P-variables can either be binary variables or multi-valued (discrete or continuous) variables. η_i represents the combined effects of unobserved characteristics (assumed to be time invariant for simplicity) and ε_{it} is the error term, assumed to be independent over time. We also assume that the interventions and control variables are uncorrelated with the error process:

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

At this stage we also assume that P and X are independent:

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

This will be relaxed in section 3. Note that equation (1) excludes spillover effects of the type where y_{it} depends on P_{jt} $(i \neq j)$ and j is not necessarily included in the sample. This point will be discussed in Section 4. In many cases (1) will represent a reduced form or "black box" regression but it can also represent a structural model.

[3] Whether P reflects an intention to treat or actual "treatment" depends on the context of the evaluation but the analysis applies to both cases.

Our interest is in the expectation (in the population) of the effect of interventions on the outcome variable. This is the expected difference $E(g_i(P_{i1}) - g_i(P_{i0}))$ which we will call the total program effect (TPE):[4]

$$\text{TPE} = E(g_i(P_{i1}) - g_i(P_{i0})).$$

Note that we do not impose a common function *g*: we allow for heterogeneity of program impact.

As an example consider a very simple special case:

$$y_{it} = \alpha_t + \beta_i P_{it} + \eta_i + \varepsilon_{it} \qquad t = 0,1 \qquad (2)$$

where P_{it} now is a binary variable rather than a vector and $P_{i0} = 0$ for all *i*. Taking first differences gives:

$$\Delta y_{i1} = \alpha + \beta_i P_{i1} + \Delta\varepsilon_{i1}$$

where $\alpha = \alpha_1 - \alpha_0$. This is analogous to the equation for a standard project evaluation, but written in differences.[5] The TPE for this case equals $E\beta_i P_{i1}$ which is related to the familiar average treatment effect on the treated (ATET)

$$\text{ATET} = \frac{\text{TPE}}{EP_{i1}}.$$

In equation (1) the terms involving the interventions and the control variables are additively separable. This allows the following identification strategy for the TPE. Assume that we have data from a random sample and that for a subsample (the "control group") there is no change in

[4] Strictly speaking this is the total effect of *changes* in the program. We use the symbol *E* for population averages and a bar over a variable for sample averages.

[5] This assumes that the autonomous trend $\alpha = \alpha_1 - \alpha_0$ is the same for all subjects (or, alternatively that the difference $\Delta\alpha_{it}$ is exogenous and can be treated as part of the residual). In the terminology of double differencing this is the assumption of parallel trends. If this assumption is questionable then data for more periods are needed to estimate how trends depend on *P*. In this paper we abstract from this complication and limit the analysis to two periods. The extension to more periods is non-trivial but conceptually straightforward.

the interventions: $P_{i1} = P_{i0}$. (At this stage we do not assume that the assignment to intended "treatment" and "control" groups is random.) Taking first differences in (1) for this group gives:

$$\Delta y_{it} = \Delta f(X_{it}) + \Delta \varepsilon_{it}.$$

This allows us to estimate $\hat{f}(X_{i1}) - \hat{f}(X_{i0})$, so that the TPE can be estimated as

$$\overline{\Delta g(P_{it})} \approx \overline{\Delta y_{it}} - \overline{\Delta \hat{f}(X_{it})}.$$

In the context we have in mind (a program consisting of multiple interventions) there will usually not be a sufficiently large control group to make this identification strategy realistic. Indeed, typically the control group will be empty: all *i* will have experienced a change in at least some components of the vector ΔP_{it}.

For the more general case we need to make a strong assumption on the functional form of $f(X)$. We will assume linearity (and suppress the subscript *t* when taking differences between the two periods considered):

$$f(X_{it}) = \gamma X_{it}.$$

Substituting this in (1) and using a first order Taylor expansion for *g(P)* gives

$$\begin{aligned} y_{it} &\approx \gamma X_{it} + g_i(P_i^*) + \nabla_P g_i(P_i^*)(P_{it} - P_i^*) + \eta_i + \varepsilon_{it} \\ &= \gamma X_{it} + \beta_i P_{it} + \eta_i^* + \varepsilon_{it} \end{aligned}$$

so that, approximately[6]

$$\Delta y_i = \gamma \Delta X_i + \beta_i \Delta P_i + \Delta \varepsilon_i. \tag{3}$$

In this case

$$\text{TPE} = E\beta_i \Delta P_i. \tag{4}$$

Note that the TPE is a weighted sum of the β_i parameters where the actual distribution of interventions provides the weights.[7]

[6] In an earlier version of this paper, Elbers and Gunning (2009), we derived this equation under much more restrictive conditions.

In general the parameters β_i will be correlated with the *P* and *X* variables.

Equation (3) can be rewritten as

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

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

The term $E(\beta_i \mid \Delta X_i, \Delta P_i)$ can be approximated linearly:[8]

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

Substitution in (4) and collecting terms gives

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

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 $T_i = E(\beta_i \Delta P_i \mid \Delta X_i, \Delta P_i)$.

Equation (5) can be estimated using the sample data. The estimated coefficients can then be used to estimate T_i as

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

The TPE can now 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 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 program effect. Note that the estimated TPE is linear in the $\hat{\alpha}$ parameters so that its standard error can be obtained straightforwardly from the covariance matrix of the OLS-coefficients.

---

[7] Note that in equation (3) $\beta_i = \nabla_P g_i(P_i^*)$. Interactions of program components are therefore one reason for treatment heterogeneity.

[8] Higher order approximations would not change the argument.

It is instructive to return to the special case of equation (2) where ΔP_i is a binary variable. In differences:

$$\Delta y_i = \alpha + \beta_i \Delta P_i + \Delta \varepsilon_i$$

In this case the quadratic approximation of $E(\Delta y_i \mid \Delta P_i)$ is exact (and in fact linear):

$$E(\beta_i \mid \Delta P_i) = \delta_0 + \Delta P_i \delta_1 = (1 - \Delta P_i) E(\beta_i \mid \Delta P_i = 0) + \Delta P_i E(\beta_i \mid \Delta P_i = 1)$$

Substitution in the regression equation gives

$$E(\Delta y_i \mid \Delta P_i) = \alpha + E(\beta_i \mid \Delta P_i = 1) \Delta P_i \qquad (7)$$

so that an OLS regression of Δy_i on ΔP_i gives an unbiased estimate of the $E(\beta_i \mid \Delta P_i = 1)$.

When can RCTs be used to estimate the TPE?

In level form (3) can be written as

$$y_{it} = \gamma X_{it} + \beta_i P_{it} + \eta_i^* + \varepsilon_{it}$$

This equation allows for two types of selection effects: P_{it} may be correlated with β_i or with the unobserved characteristics η_i^*. A correlation of P_{it} and η_i^* is dealt with by differencing, as in (3).[9] However, the TPE measures the effect of the program *inclusive* of selectivity in the assignment of program interventions resulting in a correlation of β_i and ΔP_i. This is appropriate since the way the program was assigned (in an *ex post* evaluation) or will be assigned (in an *ex ante* evaluation) is one of its characteristics. If the program was successful in part because program officers made sure the program interventions were assigned to households or locations where they expected a high impact, then the evaluation should reflect this. In fact the evaluation would be misleading if it tried to "correct" for such selection effects by presenting (if this were feasible) an estimate ($E\beta_i$) of the program's impact if it had been assigned randomly.

[9] Differencing is sufficient because of our assumption of parallel trends (cf. footnote 5).

Recall from (4) that in the linear case

$$\text{TPE} = E\beta_i \Delta P_i.$$

Clearly, if ΔP_i and β_i are independent this simplifies to

$$\text{TPE} = E\beta_i . E\Delta P_i. \qquad (8)$$

Under these assumptions $E\beta_i$ is also the average treatment effect on the treated (ATET) which in much of the project evaluation literature is the parameter of interest.[10] In this case the TPE can be estimated on the basis of an RCT, using (8): the trial would give an estimate of $E\beta_i$ and administrative or sample data could be used to estimate ΔP_i. Note that (8) can be used in two special cases. The first case is that of treatment homogeneity ($\beta_i = \beta$ for all i), the second one that of universal treatment ($P_i = 1$ for all i).[11]

When ΔP_i and β_i are not independent the ATET as established by an RCT is not a relevant parameter and estimating the TPE on the basis of RCTs can become problematic. We return to this issue in section 5.

3. Correlation between *P* and *X*

In the previous section we assumed *P* and *X* to be independent. (*P, X*) correlations are often important in evaluations. For example, changes in teacher training may induce changes in parental input. [12][13] Not all such inputs will be observed (e.g. additional parental help with

[10] But note that in project evaluations the policy variable is usually a binary variable.
[11] Imbens (2010) describes a reduction in class size in *all* California schools. This is an example of universal treatment.
[12] Deaton (2010) gives the example where random assignments made by the central government (e.g. the Ministry of Education) are partly offset by induced changes in allocations by local or provincial governments. Ravallion (2012) gives a similar example and Chen *et al.* (2009) quantify such a spillover effect in China. Similarly, the

homework will probably not be recorded); P_{it} will then be correlated with β_i and this we have already considered in the previous section. Conversely, if the parental input is observed then P_{it} will be correlated with X_{it}. In that case the TPE identifies the direct effect of *P*, but not its total effect (including the indirect effect through induced changes in *X*). If the induced effect is to be included then the affected components of ΔX_i should be omitted from the regression (5).

If causality is in the reverse direction, from ΔX_i to ΔP_i, then there is no need to amend the section 2 estimate of the TPE since there is no induced change in ΔX_i. (The asymmetry arises because in either case we are interested in the impact of changes in ΔP_i, rather than in the impact of changes in ΔX_i.)

In the general case where the direction of causality is not known it will usually not be possible to estimate the indirect effect of the program. Occasionally, however, appropriate instruments can be found so that the impact of ΔP_i on ΔX_i can be identified.

4. Spillover effects

Recall that in Section 2 we excluded spillover effects: in equation (1) y_i of case *i* does not depend on P_j of case *j*. In evaluations there are two important situations where this assumption is untenable. First, Chen *et al.* (2009) and Deaton (2010) discuss the possibility that policy in control villages is partly determined by policies in treatment villages so that the SUTVA (stable unit treatment value assumption) is violated. Indeed, if policies thus affected are not represented

political economy may be such that the central government is unable to prevent allocations being diverted to favored ethnic or political groups. In either case P_i might be correlated with β_i.

[13] This is similar to the case considered by Das *et al.* (2004, 2007) where teacher absenteeism as a result of HIV/AIDS induces greater parental input.

in the policy vector P_i this creates a classical case of omitted variable bias. In Chen *et al.* the problem arises because the data record participation in a particular program as a binary P_i variable, while other programs which may affect the outcome are initially ignored. In the approach advocated in the present paper all potentially relevant programs would in principle be included in P_i so that the problem of SUTVA violation is avoided.[14] Secondly, policies in village *j* may affect outcomes in village *i.* For example, a program aimed at an infectious disease in village *j* may affect health outcomes in the "untreated" village *i.*[15] If the external effects of policy are general equilibrium effects such as regional wage increases, it will be hard to identify the full impact of a policy. But often more structure can be imposed, e.g. by including a proxy for relevant policies in neighboring villages in the outcome regression, so that equation (3) is extended to

$$\Delta y_i = \beta_i \Delta P_i + \gamma \Delta X_i + \delta \Delta K_i + \Delta \varepsilon_i .$$

where ΔK_i is the proxy for policy changes in the neighborhood. If there is sufficient variation in K_i then δ is identified in this regression. The TPE would be $E\beta_i \Delta P_i + \delta E \Delta K_i$.

5. Regression Methods and RCTs Compared

In section 2 we showed how the TPE can be estimated using regression methods (double differencing). A natural question is whether the TPE can also be estimated using RCTs. In the Introduction we noted that using RCTs may be difficult, e.g. because in programs the dichotomy of treatment and control groups typically breaks down. However, there may be problems even in the case of binary treatments, namely under treatment heterogeneity when the probability of treatment is correlated with the individual impact parameters β_i and unknown to the evaluator.

[14] Recall that our approach does not involve a distinction between treatment and control groups: most if not all subjects receive some treatment.

[15] This has implications for sampling: since data on policies in neighboring villages are required one must sample groups (possibly pairs) of adjacent villages.

If this correlation arises through self-selection there is no problem. If the correlation arises because the policy maker targets on observables then an RCT would have to mimic this assignment, possibly by stratifying the sample on the basis of the targeting variables.

But in many government and NGO programs the "policy maker" does not directly control the *P* variables: assignment is decided by lower level staff ("program officers") on the basis of private information, variables that cannot be observed by the policy maker or the evaluator. In this case an RCT can still identify the TPE, but at the cost of having to randomize at a higher level than the treatment under consideration: randomization would apply to program officers rather than beneficiaries. This implies that the power of the statistical analysis may be reduced. It also involves losing the direct link with the intervention.

This may be illustrated with an example. Consider the following model

$$y_i = \alpha + \beta_i P_i + \varepsilon_i$$

where β_i and ε_i are independent, P_i is binary and $E\varepsilon_i = 0$. For simplicity we will consider β_i as intention-to-treat impact, so that a subject i's refusal to undergo offered treatment P_i is reflected in β_i, rather than in P_i. Program implementation involves program officers who have imperfect knowledge of β_i: they perceive $\omega_i = \beta_i + \eta_i$ and will assign treatment if and only if $\omega_i > 0$. We further assume that η_i has mean zero and is independent of β_i and ε_i. Crucially, this knowledge of program implementers is unknown to the evaluator. Denote the CDF of η_i by F. With this assignment rule P_i is exogenous (*i.e.* independent of ε_{ij}).

An RCT evaluation might involve drawing a random sample from the population and within this sample assign treatment randomly. The researcher would then estimate the program's intention to treat effect (ITE) as $E\beta_i$. The TPE would be estimated as $E\beta_i EP_i$.

This would be incorrect since, under the assumptions made above we have

$$\text{TPE} = E\beta_i P_i = E(\beta_i \mid \beta_i + \eta_i > 0)P\{\beta_i + \eta_i > 0\} = E[(1 - F(-\beta_i))\beta_i] \neq EP_i E\beta_i.$$

(Note that $E(1 - F(-\beta_i)) = P\{\beta_i + \eta_i > 0\} = EP_i$. As before, the ATET = TPE / EP_i.) The problem arises because in this case the RCT design does not mimic the actual assignment process. To obtain an unbiased estimate of the TPE randomization would have to take place at a higher level, that of the program officers.[16] The control group then consist of program officers who never "treat" and the treatment group of program officers who sometimes (but not always) treat.

The regression method we propose would lead to an unbiased estimator of the TPE using observational data for (y_i, P_i) from a random sample of the population, as shown in (7). The difference is that while the RCT approach compares average outcomes at the level of program officers the regression approach does so at the level of beneficiaries. The RCT approach therefore has lower statistical power.[17]

Moving beyond the example there is a more fundamental objection to the RCT approach if outcomes depend not only on *P* but also on *X*, as in (1). If the RCT involved randomization over actual program officers then it is unlikely that randomization can also be achieved in terms of all

[16] Duflo *et al.* (2008, pp. 3935-37) make this point in a similar context (partial compliance) concluding that "One must compare *all* those initially allocated to the treatment group to *all* those initially randomized to the comparison group".
[17] This is shown in the Supplemental Material.

the confounding X variables since program officers will not have been posted randomly across space. This introduces a correlation between X and characteristics of the program officers and hence a correlation between P and X. The two groups of program officers ("treatment" and "control") will therefore differ systematically so that internal validity is lost. Our proposed approach, by contrast, collects data at the level of beneficiaries and can therefore control for differences in X.

In summary, estimating the TPE on the basis of group averages from RCTs becomes problematic when β and P are correlated as a result of targeting on the basis of unobservables. If one randomizes at the level of beneficiaries the TPE estimator will be biased because the correlation is not taken into account. If one randomizes at the level of program officers the estimator is inefficient and, if confounders are important, may become useless.

6. An Empirical Example: Estimating the Total Program Effect for a Rural Water Supply Program in Mozambique

In this section we illustrate the estimation of the TPE with a relatively simple example based on an evaluation of the 'One Million Initiative' in Mozambique.[18]

The Initiative aims to give one million people in rural areas access to clean drinking water and adequate sanitation by constructing new water points and providing a particular type of sanitation training (CLTS). Elbers *et al.* (2012) use panel survey data for 1600 households to analyze the health impact of this program. The survey data were collected in two rounds, in 2008 and 2010, in 80 communities. There are four groups of communities: those without any intervention, those with only a water intervention or only a sanitation intervention and those

[18] Since the purpose is simply to illustrate the method we restrict the example to the specification of section 3, i.e. we do not consider the case of section 4 where X has an effect on P. Elbers *et al.* (2012) describe the Initiative.

with both types of interventions. Since the interventions were targeted on poorer communities there are significant differences between the baseline characteristics of these four groups.

Elbers *et al.* (2012) used the survey data for a double difference regression shown in the first column in Table 1. Health status was measured by a dummy variable indicating whether any household member was affected by water-borne diseases in the 6 months preceding the interview.[19] Whether there was a water point or sanitation intervention in the household's cluster (location) is measured by dummy variables. Since switching to a new, improved water point is attractive only if the new source is close, the water point intervention dummy is interacted with the distance between the household's location and the improved water source. Controls are household size and wealth and the number of under-5 children. The results suggest a substantial and significant effect of sanitation training: it reduces the probability of being affected by 8 percentage points and accounts for 20% of the decline between the two survey rounds.[20] While the effect of sanitation training is strong, access to improved water sources has no significant effect on health. This is not really surprising since the water is often not safe at the source (even for 'improved' water sources) and there is considerable contamination of water with fecal (thermo-tolerant) bacteria between the source and the point of use, a common finding in WASH studies.

In the second regression in Table 1 we include all the interaction terms suggested by equation (5).[21] In the augmented regression the additional terms are not significant, either individually or jointly (with the single exception of the interaction of distance to the new water point and household size which is marginally significant: $p = 0.09$). The coefficient of the sanitation intervention is considerably larger in absolute terms than in the original regression,

[19] By construction the health indicator is sensitive to household size. This variable is therefore included as a control.
[20] The autonomous decline of 12 percentage points is difficult to explain. It may reflect different weather conditions or differences in methods of enumerators in the two rounds.
[21] Note that some of the interactions do not introduce a new variable since the square of a binary variable is proportional to the binary variable itself.

but this is compensated by the coefficients on the interaction terms involving the sanitation intervention.

Table 1. Determinants of water-related diseases
Mozambique, 2008-2010

Dependent variable: Change in disease prevalence at household level, 6 months recall

	Minimal regression		Augmented regression	
	Estimate	Std. Error	Estimate	Std. Error
Constant (trend)	-0.119***	0.030	-0.107***	0.032
water point intervention (wpi)	0.004	0.053	0.035	0.095
sanitation intervention (si)	-0.082*	0.042	-0.172***	0.050
household size (hhs)	0.027***	0.008	0.015	0.012
number of children under-5 (ch)	0.030	0.020	0.054*	0.027
wealth (w)	-0.023	0.031	-0.028	0.047
interaction wpi * distance to water point (wd)	-0.020	0.061	-0.309	0.248
interaction wpi * si			0.116	0.101
wd squared			0.176	0.129
interaction wd * si			0.059	0.092
interaction wpi * hhs			-0.002	0.024
interaction wd * hhs			0.031*	0.018
interaction si * hhs			0.010	0.018
interaction wpi * ch			-0.028	0.074
interaction wd * ch			-0.030	0.057
interaction si * ch			-0.004	0.052
interaction wpi * w			-0.023	0.087
interaction wd * w			0.011	0.095
interaction si * w			0.028	0.075
Adjusted R^2	0.023		0.023	

Significance codes: '***' 0.01 '**' 0.05 '*' 0.1

n= 1279, mean dependent variable = -0.163

Household fixed effects regression. Clustered standard errors

Equation (6) can be used to estimate the total program effect. Table 2 summarizes the results.[22]

[22] In this illustrative example observations have not been reweighted to undo overrepresentation of poor households in the sample. Table 2 therefore contains TPE estimates for a population that looks like the sample.

Table 2 Total Program Effect of the One Million Initiative

	Augmented Regression	Original Regression
TPE	-0.050**	-0.033
(Standard error)	(0.023)	(0.023)

Standard errors corrected for clustering.

The first column in the Table is based on equation (6). The estimated TPE indicates that of the 16 percentage points decline in disease prevalence over the two year interval 5 percentage points can be attributed to the program. For comparison, the TPE is also calculated on the basis of the first regression equation in Table 1. Using this regression the TPE is smaller (in absolute value) and not significant. Since the extra coefficients in the augmented equation are not jointly significant there is no strong reason to prefer one estimate over the other. A reason to prefer the augmented regression is that it allows for heterogeneity.[23] However, it should be noted that the two TPE-estimates are within each other's confidence intervals. Obviously, a final choice of specification would require more detailed diagnostic tests and simulations, which is beyond the scope of the present paper.

The Mozambique example shows how the TPE can be calculated, allowing for treatment heterogeneity. While in this case it is not clear that treatment heterogeneity is important, in other contexts it may well be. We would advise to calculate the TPE in both ways, as in Table 2, and to test whether the difference between the estimates is significant.

[23] An argument favoring the augmented regression would be the more flexible functional form in combination with the large number of observations. For instance, Miller (2002) concludes that "…using all the available predictors will often yield predictions with a smaller MSEP [mean square error of prediction – authors] than any subset [of predictors]."

7. Conclusion

Policy makers in developing countries, NGOs and donor agencies are under increasing pressure to demonstrate the effectiveness of their program activities. At the same time there is a growing interest in using randomized controlled trials (RCTs) for impact evaluation of projects. This raises the question to what extent RCTs can be used to evaluate programs, for instance by aggregating the impact of the components of the program. This question is particularly relevant for the evaluation of budget support which is used to finance a wide variety of different activities.

The strength of RCTs is in establishing proof of principle. Going further and using RCTs to estimate the impact of programs is possible in special cases but becomes problematic if the probability of assignment is correlated with the effectiveness of the intervention, for example if teachers tend to give more attention to pupils who in their perception can benefit more from it. An RCT which randomizes at the level of beneficiaries would produce a biased estimate of the program effect since such a correlation between assignment and treatment effects would not be taken into account. Alternatively, if one randomizes at a higher level ("program officers") then the estimator is inefficient and, if confounders are important and correlated with characteristics of the program officers, it could be severely biased.

The approach proposed in this paper requires observational panel data for a representative sample of beneficiaries rather than experimental data for randomly selected treatment and control groups. If treatment is exogenous this will correctly reflect the assignment process even under treatment heterogeneity. Instead of estimating average impact coefficients for each of the various interventions of the program, we estimate the expected value (across beneficiaries) of

the total impact of the combined interventions. This parameter we have termed the total program effect (TPE). We have shown how and under what conditions regression techniques can be used to estimate the TPE in the presence of selection effects. As an example we presented TPE estimates for a rural water and sanitation program in Mozambique.

The approach has three advantages. First, by using observational data for a random sample from the population of intended beneficiaries external validity is ensured (except for general equilibrium effects). While the disadvantages of observational data are well known, this is an important advantage. Secondly, by focusing on the combined effect of program components the components are automatically correctly weighted. Finally, it avoids the problems which RCTs encounter when (as is plausible in development programs) assignment is imperfectly controlled and correlated with unobservables.

References

Banerjee, Abhijit V. and Esther Duflo (2009), 'The Experimental Approach to Development Economics', *Annual Review of Economics*, vol. 1, pp. 151-178.

Bracht, Glenn H. and Glass, Gene V. (1968), 'The External Validity of Experiments', *American Education Research Journal,* vol. 5, pp. 437-474.

Chen, Shaohua, Ren Mu, and Martin Ravallion (2009), 'Are There Lasting Impacts of Aid to Poor Areas?', *Journal of Public Economics*, vol. 93, pp. 512-528.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan (2004), 'When Can School Inputs Improve Test Scores?', Policy Research Working Paper, World Bank.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan (2007), 'Teacher Shocks and Student Learning: Evidence from Zambia', *Journal of Human Resources*, vol. 42, pp. 820-862.

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

De Kemp, Anthonie, Jörg Faust and Stefan Leiderer (2011), *Between High Expectations and Reality: an Evaluation of Budget Support in Zambia*, Bonn/The Hague/ Stockholm: BMZ/Ministry of Foreign Affairs/Sida.

Duflo, Esther, Rachel Glennerster and Michael Kremer (2008), 'Using Randomization in Development Economics Research: a Toolkit', in T. Paul Schultz and John Strauss (eds.), *Handbook of Development Economics,* Amsterdam: North-Holland, pp. 3895-3962.

Elbers, Chris and Jan Willem Gunning (2009), 'Evaluation of Development Policy: Treatment versus Program Effects', Tinbergen Institution Discussion Paper 2009-073/2.

Elbers, Chris, Jan Willem Gunning and Kobus de Hoop (2009), 'Assessing Sector-Wide Programs with Statistical Impact Evaluation: a Methodological Proposal', *World Development*, vol. 37, 2009, pp. 513-520.

Elbers, Chris, Samuel Godfrey, Jan Willem Gunning, Matteus van der Velden and Melinda Vigh (2012), 'Effectiveness of Large Scale Water and Sanitation Interventions: the *One Million Initiative* in Mozambique', Tinbergen Institute Discussion Paper 2012-069/2.

Heckman, James J., Sergio Urzua and Edward J. Vytlacil (2008), 'Understanding Instrumental Variables with Essential Heterogeneity', *Review of Economics and Statistics*, vol. 88, pp. 389-432.

Heckman James J. (1997), 'Instrumental Variables: a Study of Implicit Behavioral Assumptions Used in Making Program Evaluations', *Journal of Human Resources*, vol. 32, pp. 441-462.

Imbens, Guido W. and Joshua D. Angrist (1994), 'Identification and Estimation of Local Average Treatment Effects', *Econometrica*, vol. 62, pp. 467-476.

Ravallion, Martin (2009), 'Evaluation in the Practice of Development', *World Bank Research Observer*, vol. 24, pp. 29-53.

Ravallion, Martin (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*, vol. 50, pp. 103-114.

Rodrik, Dani (2008), 'The New Development Economics: We Shall Experiment But How Shall We Learn?', John F. Kennedy School of Government, Harvard University, HKS Working Paper RWP 08-055.

White, Howard (2006), *Impact Evaluation: the Experience of the Independent Evaluation Group of the World Bank*. Washington, DC: World Bank.

Supplemental Material

Precision of TPE estimators when treatment is exogenous but not fully controlled[24]

Using RCTs

"Program Officers" (POs) are divided into treatment- and control-POs. All subjects within the catchment area of a treatment-PO are considered as treated (i.e., we want to estimate the intention to treat effect).

Consider the following model linking outcome y_{ij} to (actual) treatment P_{ij}:

$$y_{ij} = \alpha_i + \beta_{ij} P_{ij} + \varepsilon_{ij},$$

where i refers to the program officer responsible for administrating treatment to subject j who falls within the catchment area of i. The disturbance ε_{ij} is assumed to be homoscedastic and independent of α_i, β_{ij} and P_{ij}. To model clustering by POs an officer random effect α_i is included in the model. Random effects are assumed to be i.i.d. and independent of β_{ij} and P_{ij}. We further assume that the number of subjects per PO is constant to avoid trivial complications of weighing.

The evaluator wants to estimate TPE $= E\beta_{ij} P_{ij}$ and in order to capture any selectivity in application of treatment by the program officers (PO) a random sample of POs has been drawn and subsequently been randomly divided into a group T of treatment-POs who are supposed to apply treatment to the ultimate beneficiaries j and a group C of control-POs who are asked not to give treatment to subjects. Within the catchment area of sampled POs a random sample of

[24] The context is that of section 5 in the main text of the paper.

subjects is drawn for whom we observe (at least) y_{ij}. This allows estimation of the TPE as the difference in average outcomes between group T and group C subjects: hat over TPE?

$$\hat{TPE} = \bar{y}_T - \bar{y}_C = \bar{\alpha}_T - \bar{\alpha}_C + \overline{[\beta_{ij} P_{ij}]}_T + \bar{\varepsilon}_T - \bar{\varepsilon}_C, \qquad \text{(A.1)}$$

where the bars denote sample averages over the two groups of subjects. Since this estimator is unbiased, its precision can be determined by the variance:

$$\text{MSE}(\hat{TPE}) = \left(\frac{1}{n_T} + \frac{1}{n_C} \right) \sigma_\alpha^2 + \frac{1}{N_T} [\text{var}(\beta_{ij} P_{ij})]_T + (\frac{1}{N_T} + \frac{1}{N_C}) \sigma_\varepsilon^2$$

where n_T and n_C denote the number of sampled treatment-POs and control-POs, N_T the total number of sampled subjects associated with treatment-POs, and N_C the number of sampled subjects falling under control-POs.

Regression using observational data

Now consider sampling directly at the level of subjects. Typically such a sample will also be clustered, albeit not necessarily by PO. To create a 'level playing field' we will assume that the sample has $n = n_T + n_C$ clusters with a total of $N = N_T + N_C$ subjects. For each sampled subject j from cluster i we observe P_{ij} (actual treatment) and y_{ij}. The estimator for the TPE reduces to

$$\hat{TPE} = \bar{y} - \frac{\overline{y_{ij}(1-P_{ij})}}{1-\overline{P_{ij}}} = \overline{\beta_{ij} P_{ij}} + \underline{\bar{\alpha}_i} - \frac{\overline{\alpha_i (1-P_{ij})}}{1-\overline{P_{ij}}} + \underline{\bar{\varepsilon}_{ij}} - \frac{\overline{\varepsilon_{ij}(1-P_{ij})}}{1-\overline{P_{ij}}}. \qquad \text{(A.2)}$$

Assuming as in the RCT setup that α_i is independent of P_{ij} and β_{ij} this estimator is again unbiased[25] and

$$\text{MSE}(\hat{TPE}) = \frac{1}{N} \text{var}(\beta_{ij} P_{ij}) + \text{var}\left(\underline{\bar{\alpha}_i} - \frac{\overline{\alpha_i (1-P_{ij})}}{1-\overline{P_{ij}}} \right) + \text{var}\left(\underline{\bar{\varepsilon}_{ij}} - \frac{\overline{\varepsilon_{ij}(1-P_{ij})}}{1-\overline{P_{ij}}} \right).$$

[25] Correlation of α_i and β_{ij}, P_{ij} would reflect level effects which, as explained in section 2, should be neutralized by using differenced data.

Using the delta method and the equality $E(P_{ij}-\overline{P_{ij}})^2=\frac{N-1}{N}EP(1-EP)$ it can be verified that[26]

$$\text{var}\left(\overline{\varepsilon_{ij}}-\overline{\frac{\varepsilon_{ij}(1-P_{ij})}{1-\overline{P_{ij}}}}\right)=\text{var}\left(\frac{\frac{1}{N}\sum_{ij}\varepsilon_{ij}(P_{ij}-\overline{P_{ij}})}{1-\overline{P_{ij}}}\right)\approx\frac{\overline{P_{ij}}}{1-\overline{P_{ij}}}\frac{1}{N-1}\sigma_\varepsilon^2,$$

and likewise that

$$\text{var}\left(\overline{\alpha_i}-\overline{\frac{\alpha_i(1-P_{ij})}{1-\overline{P_{ij}}}}\right)=\text{var}\left(\frac{\frac{1}{N}\sum_{ij}\alpha_i(P_{ij}-\overline{P})}{1-\overline{P_{ij}}}\right)\approx\frac{\overline{P_{ij}}}{1-\overline{P_{ij}}}\frac{1}{N-1}\sigma_\alpha^2.$$

It follows that in the regression setup precision is of order $\sqrt{N}$ while in the RCT setup precision if at best of order $\sqrt{N_T/2}$ and, if clustering of data is an issue, of order $\sqrt{n_T/2}$. (Note that if the two groups are of equal size: $N_T = N/2$, then the regression setup is twice as precise as the RCT setup.)

Covariates

Both methods fail if ε and βP are correlated. What if there are observables X_{ij} determining both P and y? This could be the result of program targeting. In that case formulas (A.1) and (A.2) can no longer be used. To account for the confounding effect of covariates a regression approach is required, also with an RCT setup. For RCTs using intention to treat by PO for estimating the TPE, efficient estimation would amount to a regression equation like

$$y_{ij}=\alpha_i+\text{TPE}I_{\{i\in T\}}+\gamma X_{ij}+\varepsilon_{ij}.$$

The reason formula (A.1) can no longer be used is that randomization over POs does not guarantee randomization over observables x_{ij}. Applying formula (A.1) we would find

$$\hat{TPE}=\overline{y}_T-\overline{y}_C=\overline{\alpha}_T-\overline{\alpha}_C+\gamma(\overline{X}_T-\overline{X}_C)+\overline{[\beta_{ij}P_{ij}]}_T+\overline{\varepsilon}_T-\overline{\varepsilon}_C.$$

[26] In this case *E* denotes an average over all possible samples.

The bias $\gamma(\overline{X}_T - \overline{X}_C)$ would vanish if $\overline{X}_T \approx \overline{X}_C$, i.e., when X_{ij} and $I_{\{i \in T\}}$. are uncorrelated.