Evaluation of Development Policy:

Treatment versus Program Effects

Chris Elbers and Jan Willem Gunning

VU University Amsterdam and Tinbergen Institute

Revised July 7, 2010

Abstract

There is a growing interest, notably in development economics, in extending project evaluation methods to the evaluation of multiple interventions ("programs"). In program evaluations one is interested in the aggregate impact of a program. In general this cannot be estimated from randomized controlled trials for individual interventions. We show how regression techniques can be used to estimate the aggregate impact with data from a representative sample of program beneficiaries. We propose a measure of program impact, the total program effect (TPE) which can be used for multiple, discrete or continuous, interventions in the presence of treatment heterogeneity. The TPE is a generalization of the treatment effect on the treated (ATET). As an example we present an estimate of the TPE for a rural water supply program in Egypt.

JEL Codes: C21, C33, O22

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

Evaluation of Development Policy:

Treatment versus Program Effects

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, increasingly governments, NGOs and donor agencies are interested in evaluating the effect of a program consisting of heterogeneous interventions such as sector-wide health and education programs. A dichotomous distinction between treatment and control groups is then 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.

The impact of the 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, Deaton, 2008, Ravallion, 2009, Rodrik, 2009, Imbens 2009, Banerjee and Duflo, 2009) even if the program involved only a single intervention. In addition, if the interventions are heterogeneous it is not even clear how one would aggregate the results of various 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: “villages”) and collecting data on the combination of interventions experienced by each village and other possible determinants of the outcome variables of interest. (There is no sharp distinction between ‘treatment’ and ‘control’ groups of beneficiaries: all beneficiaries may have been affected by the program to some degree.)

Regression techniques can then be used to estimate the impact of the various interventions.[1] 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. Similarly, 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 *et al.*, 2006) 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 is a generalization of the average treatment effect on the treated (ATET). In section 3 we show how the TPE can be estimated. Correlation between program variables and the controls is considered in section 4. In Section 5 we discuss spillovers. We illustrate the approach in Section 6 by estimating the TPE for a rural water supply program in Egypt. Section 7 concludes.

2. Impact evaluation and selection effects

Consider the following model:

[1] This approach is discussed in World Bank (2006) and Elbers *et al.* (2009). There seems to be no alternative to regression methods in this context.

$$y_{it} = c_t + P_{it}\beta_i + X_{it}\gamma + \lambda_i + \varepsilon_{it} \tag{1}$$

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 cases ('villages') sampled randomly from the population of interest. The P-variables measure the interventions to be evaluated. They can either be binary variables or multi-valued (discrete or continuous) variables. c_t denotes a time fixed effect, X observed other determinants of y, λ_i represents the combined effects of unobserved characteristics (assumed to be time invariant for simplicity) and ε is the error term, assumed independent over time. Since we allow for treatment heterogeneity the coefficients β_i are case-specific.

We will use the term *project* evaluation for the special case when there is only a single, binary P-variable, with the value 0 if i belongs to the control group and 1 for the treatment group and if there are no covariates. If P is multi-valued or if there are multiple P-variables or if outcomes depend on covariates X we will refer to the intervention as a *program*.

We assume that P, X and β are not correlated with the error term: $P_{it}, X_{it}, \beta_i \perp \varepsilon_{it}$. However, we allow for two types of selection effects: P_{it} may be correlated with β_i and with the unobserved case characteristics λ_i. Initially we assume that P and X are not correlated. This assumption will be relaxed in section 4. Note that equation (1) excludes spillover effects of the type where y_i depends on P_j ,($i \neq j$, where j is not necessarily included in the sample). This point will be discussed in Section 5.

Consider first the case of treatment homogeneity: $\beta_i = \overline{\beta}$, all i (in the population). If treatment and controls are exogenous ($P_{it}, X_{it} \perp u_{it} = \lambda_i + \varepsilon_{it}$), as in a randomized control trial (RCT), then

OLS estimation of (1) will produce an unbiased estimate of $\overline{\beta}$, which in this case clearly is the parameter of interest. In the special case of project impact evaluation $\overline{\beta}$ measures the average treatment effect on the treated (ATET) which then equals the average treatment effect (ATE).

If treatment is endogenous in the sense of "selection on the level", i.e. if P_{it} or X_{it} is correlated with the unobserved case characteristics λ_i, the equation can be estimated in first differences:

$$\Delta y_i = \alpha + \Delta P_i \overline{\beta} + \Delta X_i \gamma + \Delta \varepsilon_i, \qquad (2)$$

where $\alpha = c_1 - c_0$. Since differencing eliminates the source of the endogeneity, OLS estimation of (2) will produce an unbiased estimate of $\overline{\beta}$.

Next consider the case of treatment heterogeneity: the coefficients β_i differ across cases.[2] The differenced equation now reads:

$$\begin{aligned}\Delta y_i &= \alpha + \Delta P_i \beta_i + \Delta X_i \gamma + \Delta \varepsilon_i \\ &= \alpha + \Delta P_i \overline{\beta} + \Delta X_i \gamma + \Delta P_i (\beta_i - \overline{\beta}) + \Delta \varepsilon_i \\ &= \alpha + \Delta P_i \overline{\beta} + \Delta X_i \gamma + u_i.\end{aligned} \qquad (3)$$

The coefficients of the last equation can be estimated with OLS if ($\Delta P_i, \Delta X_i$) are not correlated with the β_i. Suppose, however, that there is "selection on the gain", because of self-selection (for example, those with high impact effects β_i choose to participate), because program staff choose the values of P_{it} on the basis of β_i or because those who expect to be assigned treatment change their behavior in response. This is the case of essential heterogeneity (Heckman, 1997,

[2] Clearly, heterogeneity may also affect α and γ but here we restrict the analysis to β-heterogeneity. Dealing with other types of heterogeneity requires different methods. For example, α-heterogeneity can be dealt with by differencing the equation once more ("triple differencing", as in e.g. Ravallion *et al.*, 2005).

Heckman *et al.*, 2006) where P_{it} is correlated with β_i. ΔP_i is then endogenous in (3) and the OLS estimate of $\bar{\beta}$ will, of course, be biased:

$$E\Delta y_i = \alpha + \Delta P_i \bar{\beta} + \Delta X_i \gamma + E\Delta P_i(\beta_i - \bar{\beta}) \neq \alpha + \Delta P_i \bar{\beta} + \Delta X_i \gamma.$$

Instrumentation cannot solve the problem (Heckman, 1997; Deaton, 2008): an instrument correlated with ΔP_i will also be correlated with u_i.

The literature suggests that in this case it may be possible to estimate the ATE for a subgroup. An example is the local average treatment effect (LATE), developed by Imbens and Angrist (1994). In principle one could build up a global estimate from a series of LATE estimates. However, we argue that in many cases this is not the parameter of interest.

Depending on the question the impact evaluation is supposed to address we consider three possibilities.

First, the evaluator may want to estimate the effect of a marginal change in ΔP for a randomly selected case *i*. In this case $\bar{\beta}$ is indeed the appropriate parameter. This case is rather special. It is relevant in an *ex post* evaluation if in the population assignments ΔP were in fact random (i.e. independent of β_i) in the evaluation period. Similarly, an estimate of $\bar{\beta}$ is useful *ex ante* if the policy maker (a) intends to make future assignments *P* either random or universal ($P_i = \bar{P}$ for all i) and (b) is in fact able to do so. This is the case in Imbens' (2009) example where the policy question is what the effect would be of a reduction in class size in *all* California schools.

Secondly, suppose the question was (*ex post*) what impact was achieved with a non-randomly assigned program $\Delta P.$ This is a central question in policy evaluations: tax payers and policy makers often want to know what interventions have actually achieved rather than what they could have achieved if designed differently. For instance, if the aim is to assess the impact against the counterfactual $\Delta P_i = 0$ (all *i*) then the impact on village *i* is $\beta_i \Delta P_i$ and the average impact in the population is $\beta_i E[\Delta P_i]$. We propose to call this the Total Program Effect (TPE). Note that the TPE is a weighted average where the actual distribution of policy changes provides the weights. The TPE measures the effect of the program, *inclusive* of selectivity in the placement of program interventions resulting in a correlation between ΔP_i and β_i. It is instructive to define the following weighted average of impact parameters β_i^j

$$\tilde{\beta}^j = E[\beta_i^j \Delta P_i^j] / E[\Delta P_i^j]$$

where the weights are the changes in the ΔP_i^j. If ΔP_i and β_i are correlated this weighted impact parameter will differ from the *unweighted* counterpart $E\beta_i^j = \bar{\beta}^j$. The TPE is a generalization of the average treatment effect on the treated (ATET): in the case of a project (i.e., a single, binary program variable ΔP_i) we have

$$\text{ATET}=\frac{\text{TPE}}{E\Delta P_i}.$$

The TPE is a generalization since it can be used in the case of multiple or multi-valued interventions. In the general case there is no natural denominator which can be used to transform the TPE into the β-dimension.

In an RCT the evaluator may be able to ensure that ΔP_i and β_i are independent and thereby obtain an estimate of $\bar{\beta}$. However, since in a reality ΔP_i and β_i will usually *not* be independent $\bar{\beta} \neq \tilde{\beta}$ so that the RCT result cannot be used to estimate the parameter of interest, the TPE: the unweighted average provides no guidance as to the value of the weighted average. This is another way in which external validity of RCTs can fail. Conversely, to the extent that participation in the RCT mimics real life participation in the program then, and only then, the RCT results can be used to estimate the program effect.

Finally, suppose the policy maker wants to estimate *ex ante* the impact of a program P and random or universal assignment is either not desirable or not feasible.[3] If future assignments are expected to be similar to past assignments then, again, what is required is an estimate of $E[\Delta P_i \beta_i]$, if necessary adjusted for differences in program size and scope. Note that the issue is not only whether the results of an RCT in, say, some village in Western Kenya can be generalized to a different context.[4] In addition, the issue is whether universal or random assignment is feasible or even desirable.

3. Estimation of the Total Program Effect

How can $E[\beta_i \Delta P_i]$ be estimated? If there are two sufficiently large groups in the sample with $\Delta P_i = 0$ in one group and $\Delta P_i \neq 0$ in the other then the TPE can be estimated quite simply from (3): the first group can be used to estimate the α and γ parameters and substitution of these in the equation for the second group can be used to derive an estimate of the TPE. However, for the situation we have in mind the condition is rarely satisfied. For example, there may be no school for which the provision of school books was the same from one year to another.

[3] Deaton (2008) 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. Similarly, the 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.

[4] See Deaton (2008) on the external validity of RCTs.

For the general case take conditional expectations in equation (3):[5]

$$E[\Delta y_i \mid \Delta P_i, \Delta X_i] = \alpha + \Delta P_i E[\beta_i \mid \Delta P_i, \Delta X_i] + \Delta X_i \gamma$$

and use a linear approximation for the conditional expectation of β_i : [6]

$$E[\beta_i^j \mid \Delta P_i, \Delta X_i] \approx \delta_0^j + \sum_k \delta_{1k}^j \Delta P_i^k + \sum_\ell \delta_{2\ell}^j \Delta X_i^\ell .$$

This gives

$$E[\Delta y_i \mid \Delta P_i, \Delta X_i] \approx \alpha + \Delta X_i^T \gamma + \sum_j \delta_0^j \Delta P_i^j + \sum_{j,k} \delta_{1k}^j \Delta P_i^k \Delta P_i^j + \sum_{j,\ell} \delta_{2\ell}^j \Delta P_i^j \Delta X_i^\ell \qquad (4)$$

Hence one can regress Δy_i on ΔP_i, ΔX_i and the interaction terms[7] of ΔP_i with ΔP_i and ΔX_i and use the estimated coefficients $\hat{\delta}_0^j, \hat{\delta}_{1k}^j, \hat{\delta}_{2l}^j$ to estimate the total program effect as

$$\text{TPE} = E[\beta_i \Delta P_i] \approx \sum_j \hat{\delta}_0^j \overline{\Delta P_i^j} + \sum_{j \le k} \hat{\delta}_{1k}^j \overline{\Delta P_i^k \Delta P_i^j} + \sum_{j,\ell} \hat{\delta}_{2\ell}^j \overline{\Delta P_i^j \Delta X_i^\ell} \qquad (5)$$

and the weighted averages $\tilde{\beta}^j$ as:

$$\tilde{\beta}^j \approx \frac{\hat{\delta}_0^j \overline{\Delta P_i^j} + \sum_{j \le k} \hat{\delta}_{1k}^j \overline{\Delta P_i^k \Delta P_i^j} + \sum_\ell \hat{\delta}_{2\ell}^j \overline{\Delta P_i^j \Delta X_i^\ell}}{\overline{\Delta P_i^j}} .$$

where the bars denote means taken over the population of interest.

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

It is instructive to consider the special case of a project, e.g. an RCT:

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

[5] Here we condition on differences. Conditioning on the levels $X_{i0}, X_{i1}, P_{i0}, P_{i1}$ leads to similar results.

[6] Higher-order approximations to $E[\beta_i^j \mid \Delta P_i, \Delta X_i]$ would not affect the conclusion: one would simply include more terms in the regression of equation (4).

[7] Obviously, combining the terms $\Delta P_j^k \Delta P_k^j$ and $\Delta P_k^j \Delta P_j^k$.

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

$$E[{}_i\beta_i \mid \Delta P_i] = \delta_0 + \Delta P_i \delta_1 = \Delta P_i E[\beta_i \mid \Delta P_i = 1] + (1 - \Delta P_i) E[\beta_i \mid \Delta P_i = 0]$$

Substitution in the regression equation gives

$$E[\Delta y_i \mid \Delta P_i] = \alpha + E[\beta_i \mid \Delta P_i = 1] \Delta P_i$$

so that an OLS regression of Δy_i on ΔP_i gives an unbiased estimate of the ATET $\tilde{\beta}$.

4. Correlation between *P* and *X*

We now relax the assumption that P and X are not correlated. Das *et al.* (2004, 2007) provide an example of such correlation: in primary schools in Zambian changes in *P*, e.g. teacher absenteeism as a result of HIV/AIDS, induce changes in parental inputs. Not all such inputs will be observed (e.g. additional parental help with 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}.[8] In that case the approach of section 3 would identify the direct effect of *P*, but not its total effect (including the indirect effect through induced changes in *X*).

More generally, from (1) it follows that

$$E\Delta y_i = \alpha + E\beta_i \Delta P_i + E\gamma \Delta X_i. \qquad (5)$$

If ΔX_i is caused by ΔP_i in the sense that:

$$\Delta X_i^k = \Delta P_i \lambda^k + \mu^k + \Delta v_i^k \qquad (6)$$

[8] This correlation was ruled out in sections 2 and 3.

where Δv_i is independent of ΔP_i, then the TPE as defined in section 2 would miss the induced effect $E\sum_{j,k}\lambda_j^k\gamma^k\Delta P_i^j$. In this case Δy_i should be regressed on a quadratic function of ΔP_i but not on terms involving ΔX_i. This gives

$$E\Delta y_i = (\alpha + \Sigma_j \mu^j\gamma^j) + E\sum_j(\delta_0^j + \Sigma_k \delta_{2k}^j\mu^k + \Sigma_k \lambda_j^k\gamma^k)\Delta P_i^j + E\sum_{j,k}(\delta_{1k}^j + \sum_m \lambda_k^m\delta_{2m}^j)\Delta P_i^k\Delta P_i^j.$$

The TPE can now be estimated as

$$TPE = \sum_j \hat{A}^j\,\overline{\Delta P_i^j} + \sum_{j\le k}\hat{B}^{jk}\,\overline{\Delta P_i^k\Delta P_i^j}$$

where $A^j = \delta_0^j + \Sigma_k \delta_{2k}^j\mu^k + \Sigma_k \lambda_j^k\gamma^k$ and $B^{jk} = \delta_{1k}^j + \sum_m \lambda_k^m\delta_{2m}^j$.

It may be desirable to decompose the TPE into the direct effect of *P* and the indirect effect (via induced changes in *X*). This can be done as follows. First, estimate the TPE in the same way as in section 3, i.e. by estimating (5) using the approximation $E[\beta_i^j \mid \Delta P_i, \Delta X_i] \approx \delta_0^j + \sum_k \delta_{1k}^j\Delta P_i^k + \sum_\ell \delta_{2\ell}^j\Delta X_i^\ell$. This gives an estimate of the direct effect, $E\beta_i\Delta P_i$. According to (6) the indirect effect is

$$\sum_{j,k}\hat{\lambda}_j^k\hat{\gamma}^k\,\overline{\Delta P_i^j}.$$

An estimate of γ is already available and (6) can be estimated to obtain estimates of λ. This gives the decomposition:

$$TPE = E\beta_i\Delta P_i + \sum_{j,k}\hat{\lambda}_j^k\hat{\gamma}^k\,\overline{\Delta P_i^j}. \qquad (7)$$

If causality is in the reverse direction, from *X* to *P*, then there is no need to amend the section 3 estimate of the TPE since there is no induced change in *X*. (The asymmetry arises because in either case we are interested in the impact of changes in *P*, rather than in the impact of changes in *X*.)

In the general case where the direction of causality is not known we can still use equation (7). However, since the error term Δv_i^j in (6) will be correlated with ΔP_i λ cannot be estimated with OLS. Estimation of the program effect then requires instruments for *P* when estimating equation (6).[9]

5. Spillover effects

Recall that in Section 2 we excluded spillover effects: in equation (1) y_i in village *i* does not depend on program P_j in village *j*. In evaluations there are two important cases where this assumption is untenable. First, Deaton (2008) and Chen et al. (2009) 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 in policy vector P_i this creates a classical case of omitted variable bias. In Chen et al. the problem arises because the data set records 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.[10] 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*. 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 relevant policies in neighboring villages in the outcome regression, so that equation (1) is extended to

[9] If there are no instruments for *P* in (6) but there are instruments for *X* in the reverse relation (*P* as a linear function of *X*) then - depending on the exclusion restrictions - it may be possible to identify the λ coefficients through 2SLS.

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

$$y_i = \alpha + \beta_i P_i + \gamma K_i + \varepsilon_i,$$

where $K_i = \sum_{j \text{ close to } i} P_j$. If there is sufficient variation in K_i then γ is identified in this regression. The effect of policy P_i would be $\beta_i + \gamma K_i$.

Of the two types of spillover effects the second is the more problematic one since it usually is much more difficult to collect data on policies in villages neighboring the sample villages.

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

In this section we illustrate the estimation of the TPE with an example.[11]

We use data collected for the evaluation of an Egyptian program of rural water supply and sanitation in the governorate of Fayoum. A survey was conducted in October-December 2008 among a total of 1500 households from 150 clusters in 77 Fayoum villages. The clusters, comprising approximately 200 households, were selected randomly from two strata. (The strata differed in whether or not major works were envisaged in the period 2008-2010.) The sample is representative for the rural population of the governorate. The households were interviewed individually by enumerators using a structured questionnaire. Apart from household characteristics (roster, housing quality, assets), the questionnaire addressed their recent health situation, water-related questions (source, use, storage), sanitation and garbage disposal, hygiene practices and awareness, and their opinions about the water, sanitation and hygiene (WASH) services they experience.[12] A second round of data collection will take place later in 2010, following the completion of a major programme to improve water pressure and to increase the

[11] 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*.

[12] The survey is discussed in detail in Netherlands Ministry of Foreign Affairs (2010). The description of the survey draws on that report.

number of households connected to piped sewage systems. For this example we can use only the base line data: equation (4) will therefore be estimated in levels rather than first differences. Implicitly we therefore assume there is no selection on the level.[13]

At the time of the survey almost all households had access to piped water but only 30% were connected to a sewage system. There are numerous government interventions in water and sanitation. Ultimately these lead to policy induced differences between households through differences in hygiene training, water pressure, water quality or access to sewage systems. The survey evidence suggests that the first channel plays no role: most respondents do not even recall having received such training let alone its contents (Netherlands Ministry of Foreign Affairs, 2010).[14] The remaining three channels are considered in the regression analysis.[15]

In this example the dependent variable is *diarrhea* prevalence, the number of instances reported by the household for a two week recall period.[16] We use three *P*-variables: *pressure* (a dummy variable, 1 for households reporting that water pressure is usually "moderate" or "good"), *sewage system* (also a dummy variable, 1 for households in clusters in which at least one household is connected to a piped sewage system, and *chlorine* (a continuous variable measuring the chlorine content of tap water, in mg/liter). There are three *X*-variables: *wealth* (an index with mean unity[17]), *household size,* and *literacy* (a dummy variable indicating whether at

[13] This is not likely but in the absence of panel data the best we can do. Using cluster fixed effects regression one can get rid of selection effects at the cluster level. However, this may well eliminate part of the program effect.

[14] While hygiene training is generally considered by sector specialists as an essential intervention the Egypt evidence is, sadly, similar to what we found in other countries (Tanzania, Yemen, Mozambique).

[15] We may well err by not including interventions in other sectors, e.g. health policies. Obviously, in a fixed effects regression the restriction to the three *P*-variables considered would be more reasonable.

[16] This is obviously restrictive since the policies considered have other impacts, e.g. a reduction in water fetching time. It is straightforward to repeat the analysis for other dependent variables.

[17] The index is based on principal components analysis, using the number of consumer durables per capita, the material used for floors and the number of rooms per capita in the house. It has been standardized at mean zero and unit variance.

least one of the household members older than 15 years is able to read). Table 1 reports descriptive statistics for these variables.

Table 1: Descriptive Statistics

variable	observations	mean	standard deviation	minimum	maximum
diarrhea	1500	0.296	0.457	0	1
pressure	1319	0.471	0.499	0	1
sewage system	1500	0.433	0.496	0	1
chlorine	1460	1.842	0.374	0	2
chlorine squared	1460	3.533	0.972	0	4
household size	1500	6.072	2.939	1	25
wealth	1500	0	1.000	-0.983	6.308
literacy	1500	0.853	0.355	0	1
pressure x sewage system	1319	0.214	0.410	0	1
pressure x chlorine	1280	0.903	0.959	0	2
sewage system x chlorine	1460	0.786	0.931	0	2
pressure x household size	1319	2.836	3.594	0	20
pressure x wealth	1319	0.003	0.654	-0.983	6.308
pressure x literacy	1319	0.411	0.492	0	1
sewage system x household size	1500	2.520	3.421	0	20
sewage system x wealth	1500	0.044	0.705	-0.983	6.308
sewage system x literacy	1500	0.377	0.485	0	1
chlorine x household size	1460	11.229	6.079	0	50
chlorine x wealth	1460	-0.001	1.899	-1.965	11.735
chlorine x literacy	1460	1.567	0.741	0	2

In Table 2 we report the regression corresponding to equation (4), using all *P* and *X* variables and their interactions. If we accepted these results the best estimate of the effect of the program would be obtained by taking for each term involving a *P*-variable the mean value of the regressor multiplied by the regression coefficient and summing over terms. (Note that this defines the counterfactual as the case where all *P*-variables equal 0.) This gives an estimate of -0.108 and a t-score of -0.84: the joint effect of the three *P*-variables on diarrhea prevalence is very substantial (reducing prevalence by 11 percentage points, from 41% to 30%) but not statistically significant.

Table 2: Determinants of Diarrhea Prevalence
(OLS)

	coefficient	t-score
pressure	0.157	0.80
sewage system	-0.212	-0.96
chlorine	-0.320	-1.27
chlorine squared	0.055	0.66
household size	-0.047	-1.01
wealth	-0.147	-1.15
literacy	-0.071	-0.31
pressure x sewage system	-0.005	-0.10
pressure x chlorine	-0.091	-1.01
sewage system x chlorine	0.059	0.55
pressure x household size	-0.021	-1.79
pressure x wealth	0.024	0.69
pressure x literacy	0.123	1.55
sewage system x household size	-0.011	-0.93
sewage system x wealth	-0.002	-0.07
sewage system x literacy	0.169	2.12
chlorine x household size	0.038	1.58
chlorine x wealth	0.056	0.84
chlorine x literacy	-0.073	-0.62
constant	0.749	2.15

number of observations 1280. Adjusted $R^2 = 0.0203$

In view of the large number of insignificant coefficients we attempt to find a more parsimonious specification. We use a forward stepwise regression approach with the restriction that all *X*-variables are included. (This is to prevent omitted variable bias in the form of program variables picking up the effect of controls.). This leads to eight regressors: the controls *household size, wealth*, *literacy,* the policy variable *sewage system* and five interaction terms. The results of this regression are shown in Table 3.

Table 3: Determinants of Diarrhea Prevalence (OLS, parsimonious specification)

	Coefficient	t-score
household size	0.025	3.56
wealth	-0.026	-1.69
literacy	-0.189	-3.12
pressure x household size	-0.022	-2.88
pressure x literacy	0.104	1.91

sewage system x household size	-0.009	-0.94
sewage system x literacy	0.183	2.43
sewage system	-0.131	-1.68
constant	0.334	5.35

Forward stepwise regression. Entry condition p < 0.30 for all regressors. X-variables restricted to be included. Robust standard errors have been used to calculate t-scores.

N = 1319. $R^2 = 0.0278$.

The corresponding TPE estimate is obtained by summing the five regressors involving policy variables, weighted by the corresponding OLS-coefficients. (As before, the counterfactual is the case where all *P*-variables equal 0.) This gives the total program effect as -0.03 (a reduction of diarrhea prevalence by 3 percentage points) with a t-score of -1.89. It is worth noting that the TPE suggests a much smaller reduction in diarrhea prevalence than the naive estimate of 11 percentage points obtained from the Table 2 regression.[18]

7. Conclusion

Policy makers, 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 projects that constitute the program. This is particularly relevant for the evaluation of budget support which is used to finance a wide variety of different activities.

Unfortunately, the scope for using RCTs in this context is quite limited: since “program assignment” is typically non-random by design or necessity, effects established by an RCT are not directly relevant for population-wide programs if the impact differs across beneficiaries

[18] The 11% is outside the confidence interval of the TPE.

(treatment heterogeneity). In addition, many policy activities cannot be summarized by a binary treatment variable. For example, what matters in an education program is not just whether a school receives textbooks but also how many.

In this paper we have argued that average causal treatment effects estimated from RCTs are of limited value in program evaluation, i.e. when there are multiple or multi-valued interventions. RCTs identify the parameter of interest only if the effect is the same for all beneficiaries or if the program would be applied universally and involves no externalities. Usually, however, the interest (either *ex post* or *ex ante*) is in the effectiveness of a program where random or universal assignment (or intention to treat) is neither feasible nor desirable. The assessment of the impact of the program should reflect this.

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). Rather than estimating (unweighted) average impact coefficients for each of the various interventions making up 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). This can be estimated if one replaces the conditional expectation of the impact coefficients by an approximation (possibly linear) in intervention and control variables. We have shown how and under what conditions the TPE can be estimated in the presence of selection effects. The approach is illustrated with an example for a rural water and sanitation program in Egypt.

References

Banerjee, Abhijit V. and Esther Duflo (2008), ‘The Experimental Approach to Development Economics’, NBER Working Paper 14467.

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 (2008), ‘Instruments for Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development’, NBER Working Paper 14690.

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, 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.

Heckman, James J., Sergio Urzua and Edward J. Vytlacil (2006), ‘Understanding Instrumental Variables with Essential Heterogeneity’, NBER Working Paper 12574.

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. (2009), ‘Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)’, NBER Working Paper 14896.

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

Netherlands Ministry of Foreign Affairs (2010), *Impact Evaluation: Drinking Water Supply and Sanitation Programme Supported by the Netherlands in Fayoum Governorate, Arab Republic of Egypt, 1990 – 2009*, Policy and Operations Evaluation Department.

Ravallion, Martin, Emanuela Galasso, Teodoro Lazo, and Ernesto Philipp (2005), ‘What Can Ex-Participants Reveal about a Program's Impact?’, *Journal of Human Resources*, vol. 40, pp. 208-230.

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

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.

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