TI 2009-073/2
Tinbergen Institute Discussion Paper

Evalution of Development Policy: Treatment versus Program Effects

Chris Elbers
Jan Willem Gunning

VU University Amsterdam, and Tinbergen Institute.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

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

Evaluation of Development Policy: Treatment versus Program Effects[1]

Chris Elbers and Jan Willem Gunning

VU University Amsterdam and Tinbergen Institute

Revised August 2009

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 rather than the effect on individual beneficiaries. In many situations randomized controlled trials cannot identify this impact. We propose a measure of program impact, the total program effect (TPE), which is a generalization of the treatment effect on the treated (ATET). We show how the TPE can be estimated.

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 Jean-Marie Baland, Martin Ravallion and Vincenzo Verardis for helpful discussions and to Remco Oostendorp for very useful comments on an earlier version of the paper.

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 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 runs 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 intervention is homogeneous. In addition, if the interventions are heterogeneous it is not even clear how one would aggregate the results of various RCTs. One can, however, modify project evaluation techniques to make them suitable for a sector-wide context. This would involve drawing a

representative sample of beneficiaries (e.g. households, schools, or communities) and collecting data on the combination of interventions experienced by each beneficiary and other possible determinants of the outcome variables of interest. Regression techniques can then be used to estimate the impact of the various interventions and from this an aggregate impact of the program can be derived.[2]

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 impact heterogeneity, either through self-selection or through the allocation decisions of program officials. In either case the intervention variables would be endogenous.

If the endogeneity is due to impact heterogeneity ("selection on the gain", Heckman *et al.*, 2006) then this should be incorporated in an estimate of the program effect. (For projects this makes ATET, the average treatment effect on the treated, a much more relevant parameter than ATE, the average treatment effect in the population, as we will see below.)

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 using data representative for the

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

population of interest. Correlation between program variables and the controls is considered in section 4. Section 5 concludes.

2. Impact evaluation and selection effects

Consider the following model:

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

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 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 impact 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 3.

Consider first the case of treatment homogeneity: $\beta_i = \overline{\beta}$, all i (in the population). If treatment is 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.[3] The differenced equation now reads:

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

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

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

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). While Imbens (2009) suggests with the title of his paper

[3] Clearly, heterogeneity can also affect α and γ but here we restrict the analysis to β-heterogeneity. To take care of 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).

("LATE or Nothing") that there is no alternative to the LATE, we argue that in many cases not even the (global) ATE is 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 ΔP. This is a central question in policy evaluations: tax payers and policy makers want to know what interventions have actually achieved rather than what they could have achieved if designed differently, e.g. if targeted on a particular group. This calls for an estimate of $E[\Delta P_i \beta_i]$, which we will call the *total program effect* (TPE), the average (per case) effect of the program, *inclusive* of selectivity in the placement of program interventions or any unobserved responses by intended

beneficiaries resulting in a correlation between ΔP_i and β_i.[4] It is instructive to define the following weighted average of impact parameters β_i^j

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

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 *un*weighted counterpart $E\beta_i^j = \bar{\beta}^j$. Note that $\text{TPE} = \sum_j \tilde{\beta}^j E\Delta P_i^j$ and that in the case of a project (i.e., a single, binary program variable ΔP_i) the TPE is related to the average treatment effect on the treated $\text{ATET} = E(\beta_i \mid \Delta P_i = 1)$ as

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

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 the program 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. This is another way in which external validity 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.[5] If future

[4] If the endogenous responses are observed they result in a correlation of *P* and *X*. This is considered in section 4.

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.[6] In addition, the issue is whether universal or random assignment is feasible or even desirable.

3. Estimation of the Total Program Effect

How can $E[\Delta P_i \beta_i]$ be estimated? Take conditional expectations in equation (3):[7]

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

and use a linear approximation for the conditional expectation of β_i:[8]

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

This gives

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

---

[5] Deaton (2008) gives the example of a non-monolithic public sector 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 may be correlated with β_i.

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

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

[8] 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).

Hence one can regress Δy_i on

$\Delta P_i, \Delta X_i$ and the interaction terms of ΔP_i with ΔP_i and ΔX_i [9] and use the estimated coefficients $\hat{\delta}_0^j, \hat{\delta}_{1k}^j, \hat{\delta}_{2l}^j$ to estimate the total program effect as

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

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

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

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 $\tilde{\beta}$.

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

4. Correlation between *P* and *X*

Das *et al.* (2004, 2007) show that 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}.[10] 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

$$E\Delta y_i = \alpha + E\Delta P_i \beta_i + E\Delta X_i \gamma. \tag{5}$$

If ΔX_i is caused by ΔP_i in the sense that:

$$\Delta X_i^k = \Delta P_i \lambda^k + \mu^k + \Delta v_i^k \tag{6}$$

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

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

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

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 $A^j = \delta_0^j + \sum_k \delta_{2k}^j \mu^k + \sum_k \lambda_j^k \gamma^k$ 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, $\overline{EP_i \beta_i}$. 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:

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

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*, not 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).[11]

3. 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, notably under 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 discussed when RCT estimates can be used in program evaluation. An RCT with compulsory assignment will produce an estimate of the average treatment

[11] 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.

effect (ATE) in the population from which the RCT sample was drawn. This is also the parameter of interest for an evaluation of the effect in a larger population, provided the sample was representative for that population. This would e.g. be the case 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 is neither feasible nor desirable. The variable of interest is then the total program effect (TPE) introduced in this paper. We have shown how and under what conditions the TPE can be estimated in the presence of selection effects.

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.

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.

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.