The World Bank Economic Review Advance Access published September 7, 2013

Evaluation of Development Programs: Randomized Controlled Trials or Regressions?

Chris Elbers and Jan Willem Gunning

Can project evaluation methods be used to evaluate programs: complex interventions involving multiple activities? A program evaluation cannot be based simply on separate evaluations of its components if interactions between the activities are important. In this paper a measure is proposed, the total program effect (TPE), which is an extension of the average treatment effect on the treated (ATET). It explicitly takes into account that in the real world (with heterogeneous treatment effects) individual treatment effects and program assignment are often correlated. The TPE can also deal with the common situation in which such a correlation is the result of decisions on (intended) program participation not being taken centrally. In this context RCTs are less suitable even for the simplest interventions.

The TPE can be estimated by applying regression techniques to observational data from a representative sample from the targeted population. The approach is illustrated with an evaluation of a health insurance program in Vietnam. JEL codes: C21, C33, O22

Experimental methods 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 various interventions, e.g. sector-wide health or education programs (De Kemp *et al.*, 2011). Program evaluation faces two complications. First, a sharp 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

The authors are professors at the VU University Amsterdam and fellows of the Tinbergen Institute. Their addresses are c.t.m.elbers@vu.nl and j.w.gunning@vu.nl (corresponding author). They are grateful to Remco Oostendorp, Menno Pradhan, Martin Ravallion, Elisabeth Sadoulet, Finn Tarp, the editors and two anonymous referees of this *Review* and to seminar participants in Amsterdam, Namur, Oxford and Paris for very valuable comments on previous versions. A supplemental appendix to this article is available at http://wber.oxfordjournals.org/.

THE WORLD BANK ECONOMIC REVIEW, pp. 1–14 doi:10.1093/wber/lht025

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 clear how RCT evaluations of individual components of the program should be combined to an overall assessment of the program. However, regression techniques can be used for program evaluation. 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.[1] In this paper this approach is generalized by allowing for treatment heterogeneity and a way of estimating aggregate program impact is proposed.

Obviously, the intervention variables are likely to be endogenous in a regression analysis. For example, an unobserved variable such as the political preferences of the community may affect both the impact variable of interest and the intervention. Also, the impact of the intervention will differ between beneficiaries and the allocation of interventions across beneficiaries may be based on such treatment heterogeneity, either through self-selection or through the allocation decisions of program officers. Heckman (1997) and Heckman *et al.* (2008) call this "selection on the gain". The first complication is usually dealt with by using panel data or by randomized assignment of treatment. The second complication is much more serious. It may be particularly hard for RCTs when program assignment in practice cannot be mimicked by assignment to the treatment arm in an RCT since this would not capture the way program officers take their decisions. However, it will be shown that regression techniques can be adapted so as to produce an appropriate estimate of the program effect.

The paper is organized as follows. In the first section the total program effect (TPE) is introduced. This measure extends the average treatment effect on the treated (ATET). The TPE is suitable for complex interventions and can deal with selection on the gain (treatment heterogeneity). Then two complications are considered: correlation between program variables and the controls in section 2 and spillover effects in section 3. Section 4 investigates whether estimating the TPE using RCTs is an alternative. The approach is illustrated in section 5 by estimating the TPE for a health insurance intervention in Vietnam. Section 6 concludes.

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

I. The Total Program Effect (TPE)

Consider the following model:

$$y_{it} = \alpha X_{it} + \beta_i P_{it} + \gamma_t + \eta_i + \varepsilon_{it} \tag{1}$$

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 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.[2] The P-variables can either be binary variables or multi-valued (discrete or continuous) variables. α and β_i are vectors of parameters, γ_t denotes a time effect and η_i represents time-invariant unobserved characteristics and ε_{it} is the error term, assumed to be independent over time. It is also assumed 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 P and X are assumed to be independent:

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

This will be relaxed in section 2. 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 3. In many applications (1) will represent a reduced form or "black box" regression, but it can also represent a structural model.

The evaluator is interested in the expectation (in the population) of the effect of interventions on the outcome variable, the total program effect (TPE):[3]

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

Note that the impact parameters β_i need not be the same for all i: heterogeneity of program impact is allowed.

As an example consider a very simple special case:

$$y_{it} = \beta_i P_{it} + \gamma_t + \eta_i + \varepsilon_{it},\ t = 0, 1 \tag{2}$$

where P_{it} now is a binary variable rather than a vector, $P_{i0} = 0$ for all i and $P_i = P_{i1} - P_{i0}$. Taking first differences gives:

[2] Here P reflects "actual" treatment. In principle it could reflect "intended" treatment if intended treatment can be observed, e.g. because intended beneficiaries were offered vouchers.

[3] Strictly speaking this is the total effect of *changes* in the program. The symbol E is used for population averages and a bar over a variable for sample averages. Note that the total program effect does not include general equilibrium effects of the program.

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

where $\gamma = \gamma_1 - \gamma_0$. This is analogous to the equation for a standard project evaluation, but written in differences.[4] The TPE for this case equals $E\beta_i P_i$ which is related to the familiar average treatment effect on the treated (ATET)

$$\text{ATET} = \frac{\text{TPE}}{EP_i}.$$

In another special case of equation (1) the TPE can be identified as follows. Assume that data are available from a random sample and that for a subsample (the "control group") there is no change in the interventions: $P_{i1} = P_{i0}$. (At this stage it is not assumed that the assignment to intended "treatment" and "control" groups is random.) Taking first differences in (1) for this group gives:

$$\Delta y_i = \alpha\Delta X_i + \gamma + \Delta\varepsilon_i \text{ if } P_i = 0.$$

This allows estimation of α and hence $\hat{\alpha}\overline{\Delta X_i}$ so that the TPE can be estimated as

$$\hat{\text{TPE}} = \overline{\Delta y_i} - \hat{\alpha}\overline{\Delta X_i}.$$

However, in a program consisting of multiple interventions, the context of this paper, 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_i.

For this more general case

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

Allowing for "selection on the gain", correlation between impact parameters β_i and the program variables P_i and also for correlation between β_i and X_i equation (3) can be rewritten as

$$\Delta y_i = \alpha\Delta X_i + E(\beta_i|\Delta X_i, \Delta P_i)\Delta P_i + \gamma + \omega_i, \tag{4}$$

[4] This assumes that the autonomous trend $\gamma = \gamma_1 - \gamma_0$ is the same for all subjects (or, alternatively that the difference $\Delta\gamma_{it}$ 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. This paper abstracts from this complication and limits the analysis to two periods. The extension to more periods is non-trivial but conceptually straightforward.

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

The term $E(\beta_i|\Delta X_i, \Delta P_i)$ can be approximated linearly:[5]

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

Substitution in (4) and collecting terms gives

$$\Delta y_i = \gamma + \theta_1 \Delta X_i + \theta_2 \Delta P_i + \theta_3 \Delta X_i \otimes \Delta P_i + \theta_4 \Delta P_i \otimes \Delta P_i + \omega_i \tag{5}$$

where

$$\theta_2 \Delta P_i + \theta_3 \Delta X_i \otimes \Delta P_i + \theta_4 \Delta P_i \otimes \Delta P_i$$

is the approximation of $T_i = E(\beta_i \Delta P_i | \Delta X_i, \Delta P_i)$.

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

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

The TPE can now be estimated as the average of $\hat{T}_i$ in the sample.

$$T\hat{P}E = \frac{1}{n}\sum_i \hat{T}_i = \hat{\theta}_2 \overline{\Delta P_i} + \hat{\theta}_3 \overline{\Delta X_i \otimes \Delta P_i} + \hat{\theta}_4 \overline{\Delta P_i \otimes \Delta P_i} \tag{6}$$

where bars denote sample averages.[6]

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. Since the estimated TPE is linear in the $\hat{\theta}$ parameters its standard error can be obtained from the covariance matrix of the OLS-coefficients.

It is instructive to consider the special case of equation (5) where $D_i = \Delta P_i$ is a binary variable taking the value 1 for the treatment group and 0 for the control group, i.e. the case of a difference-in-difference analysis. Equation (5) now reduces to

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

since in this case ${D_i}^2 = D_i$. Compared to a standard diff-in-diff regression this equation contains the interaction term $D_i\, \Delta X_i$.

[5] Higher order approximations would not change the argument but it should be noted that the number of regressors expands very rapidly. De Janvry *et al.* (2012) account for treatment heterogeneity in a similar way in the context of a schooling program.

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

The total program effect will in this case be estimated as

$$\hat{\text{TPE}} = \hat{\theta}_2\overline{D_i} + \hat{\theta}_3\overline{D_i \Delta X_i} \tag{7}$$

This shows that when the sample is representative sample means can be used to construct the total program effect. The interaction term in (7) avoids the bias resulting from correlations between treatment effects and either program participation or controls.

Many diff-in-diff studies do not include the interaction terms (e.g., Khandker *et al.*, 2009 or Almeida and Galasso, 2010). Studies that do often report estimates of impact for different values of the controls X which makes it difficult to assess the aggregate impact of a program.

Equation (1) 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).[7] 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 obviously 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.

Recall that in the special, binary case of a 'project' evaluation $\text{TPE} = E\beta_i\Delta P_i = \text{ATET} \times E\Delta P_i$. If administrative data can be used to estimate $E\Delta P_i$ the question arises whether the ATET is identified in an RCT. Obviously this is the case if $\beta_i = \beta$ for all i. More generally, if ΔP_i and β_i are independent the TPE can be estimated on the basis of an RCT: the trial would give an estimate of $E\beta_i$ which in this case is also the ATET. A special case of independence is that of universal treatment ($P_i = 1$ for all i).[8] In the most general case when ΔP_i and β_i are not independent the ATET as established by an RCT may differ from the ATET in the population and estimating the TPE on the basis of RCTs can become problematic. This issue will be considered in section 4.

II. Correlation between P and X

In the previous section P and X were assumed to be independent. (P, X) correlations are often important in evaluations. For example, changes in teacher

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

[8] Imbens (2010) describes a reduction in class size in *all* California schools. This is an example of universal treatment.

training may induce changes in parental input.[9,10] 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 was 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 1 estimate of the TPE since there is no induced change in ΔX_i. (The asymmetry arises because in either case the interest is 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.

III. Spillover Effects

Recall that in section 1 spillover effects were excluded: 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 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.[11] 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.[12] If the external effects

[9] 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 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.

[10] 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.

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

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

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

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

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 then be $E\beta_i\, \Delta P_i + \delta\, E\Delta K_i$.

IV. Regression Methods and RCTs Compared

In section 1 it was shown how the TPE can be estimated using regression methods. A natural question is whether the TPE can also be estimated using RCTs. Using RCTs may be difficult, e.g. because in programs the distinction between treatment and control groups may break 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. If this correlation arises through self-selection then the usual response is to consider the average treatment effect on the treated rather than the average treatment effect in the population. If, however, the correlation arises at a higher level, e.g. 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

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

where β_i and ε_i are independent, P_i is binary and $E\varepsilon_i = 0$. For simplicity β_i will be considered as the 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 $\omega_i = \beta_i + \eta_{\mathrm{i}}$ and will assign treatment if and only if $\omega_i > 0$. Assume that η_i has mean zero and is independent of β_i and ε_i. Crucially, this knowledge of program officers 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 assign treatment randomly within this sample. 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

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

(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.[13] 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 proposed regression method gives an unbiased estimator of the TPE using observational data for (y_i, P_i) from a random sample of the population. 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.[14]

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 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.[15] The 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 $\beta$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 inconsistent.

[13] 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".

[14] This is shown in the supplemental appendix.

[15] This is shown in the supplemental appendix.

TABLE 1. Data for the Vietnam Insurance Example

Variable: change in (average)	Mean	Std. Dev	Min	Max
Arm circumference (cm)	1.154	2.013	−7.3	9.4
Height (cm)	5.175	11.35	−49.57	39.84
Body weight (kg)	2.983	6.544	−27.75	26.25
Health expenditure (‘000 Dong)	1,081	5,519	−8808	23,3965
Total consumption expenditure (‘000 Dong)	6,513	8,009	−22,988	11,6826
Insurance (binary at individual level)	0.170	0.268	0	1
School attended[16]	−0.017	0.683	−3.5	3
Currently attending school (binary at individual level)	0.082	0.388	−2	2
Gender	0.002	0.138	−0.75	1
Age	3.522	8.299	−48.43	48.6
Farm dummy	−0.079	0.421	−1	1
Household size	−0.267	1.696	−18	11

Note: The number of observations varies between 4299 and 4305.

Source: authors’ calculations using the Vietnam Living Standard Surveys 1992–3, 1997–8.

V. AN EMPIRICAL EXAMPLE: ESTIMATING THE TOTAL PROGRAM EFFECT FOR A HEALTH INSURANCE PROGRAM IN VIETNAM

To illustrate how the total program effect can deviate from a naïve approach to calculating the effect of a program a study of the impact of a health insurance program in Vietnam (Wagstaff and Pradhan, 2005) is reconsidered. Health insurance was introduced between the 1992-93 and the 1997-98 rounds of the Vietnam Household Living Standards Survey (General Statistics Office of Vietnam, 1993 and 1998). To account for possible treatment heterogeneity Wagstaff and Pradhan match households on propensity scores and then compare changes in health outcomes (as well as some non-health outcomes) between insured households or individuals and (matching) uninsured households or individuals. They find modest favorable effects on children’s nutritional status, a mild effect on health expenditure and a sizeable effect on non-health spending.

A propensity score based approach is not suitable for calculation of a total program effect since the common support requirement in a PSM approach will exclude part of the population in a systematic way. Therefore the Vietnam data are used to estimate the effect of the program using a standard diff-in-diff approach *i.e.* without allowing for heterogeneity (labeled ‘naïve’). The results are compared with an estimate of the TPE. In this case the ‘program’ is a simple intervention.[16] This makes a comparison with a standard approach clearer.

The data are summarized in Table 1. A difficulty is that some of the outcome variables are individual anthropometric measurements while only households can be matched between survey rounds. Therefore the individual measurements

[16] It should be noted that the intervention variable is not binary (as it would be in a ‘project’) since insurance enrollment is measured as an average at the household level.

Table 2. Total Program Effects

Dependent variable	Naïve program effect[†] (I)(s.e.)	Total program effect[††] (II)(s.e.)	R-squared of underlying regressions		Remarks
			I	II	
Arm circumference	.022(.029)	0.090***(0.027)	0.22	0.23	
Height	−0.190(0.154)	.095(0.139)	0.34	0.36	
Body weight	0.167*(0.083)	0.384***(0.074)	0.31	0.33	
Health expenditure	−28.08(60.59)	−52.79(51.01)	0.03	0.04	Total consumption included in controls
Health expenditure	55.41(66.42)	64.32(52.87)	0.00	0.00	Total consumption expenditure not included
Total consumption expenditure	626.7***(110.9)	888.8***(105.7)	0.10	0.12	Total consumption expenditure not included

Notes: Robust clustered standard errors in parentheses. In all but the health expenditure regressions squared intervention and interactions of controls with intervention are jointly significant. Significance: * indicates 5% threshold, *** 0.1%.

[†] The naïve program effect is calculated as the regression coefficient on the insurance variable time the estimated population mean of that variable.

[††] The total program effect is calculated according to equation (6).

The sampling errors on the estimated population means are not taken into account.

Source: authors' calculations using the Vietnam Living Standard Surveys 1992–3, 1997–8.

have been averaged per household - a crude procedure only suitable for the current purpose of illustrating the TPE. Lacking information on 1992-3 the sampling weights from 1997-8 are used; clustering is also based on the 1997-8 survey round.

The outcome variables considered are changes in arm circumference, height, body weight, health expenditure and total expenditure. The explanatory variables are the other variables shown in Table 1 (insurance status and the controls school attended, currently attending school, gender, age, a farm dummy, household size): and their interaction with the intervention variable are used as explanatory variables. When total expenditure is not a dependent variable it is also used as control variable.

Table 2 summarizes the results. First a naïve regression is run (without interaction terms) and the implied program effect (calculated as the regression coefficient of insurance times mean insurance). This naïve program effect is then compared with a TPE calculated as in equation (6).

The results show striking differences between the two methods. In the case of arm circumference the standard method would have led to the conclusion that insurance had no (significant) effect. Once treatment heterogeneity is allowed for the effect is in fact highly significant albeit very small. For height neither method finds a significant effect. For body weight both methods show a significant increase but the effect is more than twice as large when heterogeneity is allowed for.

Insurance appears to have no significant effect on health expenditure irrespective of the method used. Both methods do find a substantial (and significant) effect of insurance on total consumption. Again, the effect is stronger once one takes heterogeneity into account.

Obviously, there is no reason why these results should generalize. However, they do suggest that treatment heterogeneity can have a substantial effect on the estimates of a program's impact. A simple way to investigate this possibility is to test for the joint significance of the coefficients on the variables which would not normally be included in the regression: the interactions of treatment variables with themselves and with the controls. When this test indicates that heterogeneity may be an issue it is advisable to calculate the TPE.

VI. 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 or of NGOs which typically involve a wide variety of 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, teachers may give more attention to children who they think can benefit more from it. An RCT which randomizes at the level of beneficiaries (children) would produce a biased estimate of the program effect by ignoring this correlation between assignment and treatment effects. Alternatively, randomization at the appropriate level (teachers) would require a larger sample for the same precision. If confounders are important and correlated with characteristics of the program officers, the RCT-based estimate of the program's impact would even be inconsistent.

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, the expected value (across beneficiaries) of the total impact of the combined interventions is estimated. This gives the total program effect (TPE). The paper has shown how and under what conditions regression techniques can be used to estimate the TPE in the presence of selection effects. As an example TPE estimates for a simple intervention: a health insurance program in Vietnam were presented. The example shows that allowing for heterogeneity can lead to very different estimates of a program's effect. The proposed method offers a simple way of dealing with such heterogeneity.

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

References

Almeida, R. K., and E. Galasso. 2010. "Jump-starting Self-employment? Evidence for Welfare Participants in Argentina." *World Development* 38(5): 742–55.

Banerjee, A. V., and E. Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–78.

Bracht, G. H., and G. V. Glass. 1968. "The External Validity of Experiments." *American Education Research Journal* 5(4) : 437–74.

Chen, S., R. Mu, and M. Ravallion. 2009. "Are There Lasting Impacts of Aid to Poor Areas?" *Journal of Public Economics* 93(3): 512–28.

Das, J., S. Dercon, J. Habyarimana, and P. Krishnan. 2004. "When Can School Inputs Improve Test Scores?" World Bank Policy Research Working Paper 3217. World Bank, Washington, DC.

———. 2007. “Teacher Shocks and Student Learning: Evidence from Zambia.” *Journal of Human Resources* 42(4): 820–62.

Deaton, A. 2010. “Instruments, Randomization, and Learning about Development.” *Journal of Economic Literature* 28(2): 424–55.

De Janvry, A., F. Finan, and E. Sadoulet. 2012. “Local Electoral Incentives and Decentralized Program Performance.” *Review of Economics and Statistics* 94(3): 672–85.

De Kemp, A., J. Faust, and S. 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, E., R. Glennerster, and M. Kremer. 2008. “Using Randomization in Development Economics Research: a Toolkit.” In T. Paul Schultz and J. Strauss, eds., *Handbook of Development Economics*. Amsterdam, The Netherlands: North-Holland, 3895–962.

Elbers, C., and J. W. Gunning. 2009. “Evaluation of Development Policy: Treatment versus Program Effects.” Tinbergen Institute Discussion Paper 2009-073/2. University of Amsterdam, The Netherlands.

Elbers, C., J. W. Gunning, and K. de Hoop. 2009. “Assessing Sector-Wide Programs with Statistical Impact Evaluation: a Methodological Proposal.” *World Development* 37(2): 513–20.

General Statistics Office of Vietnam. 1993. Living Standards Survey 1992-93. http://go.worldbank.org/JZFNBLXM80, accessed May 2013.

———. 1998. Living Standards Survey 1997-98. http://go.worldbank.org/4QR0OSXMD0, accessed May 2013.

Heckman, J. J. 1997. “Instrumental Variables: a Study of Implicit Behavioral Assumptions Used in Making Program Evaluations.” *Journal of Human Resources* 32(3): 441–62.

Heckman, J. J., S. Urzua, and E. J. Vytlacil. 2008. “Understanding Instrumental Variables with Essential Heterogeneity.” *Review of Economics and Statistics* 88(3): 389–432.

Imbens, G. W., and J. D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62(2): 467–76.

Khandker, S. R., Z. Bakht, and G. B. Koolwal. 2009. “The Poverty Impact of Rural Roads: Evidence from Bangladesh.” *Economic Development and Cultural Change* 57(4): 685–722.

Ravallion, M. 2009. “Evaluation in the Practice of Development.” *World Bank Research Observer* 24(1): 29–53.

———. 2012. “Fighting Poverty One Experiment at a Time: a Review of Abhijit Banerjee and Esther Duflo’s *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*.” *Journal of Economic Literature* 50(1): 103–14.

Rodrik, D. 2008. “The New Development Economics: We Shall Experiment But How Shall We Learn?” HKS Working Paper RWP 08-055. John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Wagstaff, A., and M. Pradhan. 2005. “Health Insurance Impacts on Health and Nonmedical Consumption in a Developing Country.” World Bank Policy Research Working Paper 3563. World Bank, Washington, DC.

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