

*METHODOLOGY FOR EVALUATIONS OF BUDGET
SUPPORT OPERATIONS AT COUNTRY LEVEL*

Tools for "Step 2":
The evaluation of the impact of government
strategies

April 2009

Preface

This paper is based on (and copies extensively from) the discussion paper by Jan Willem Gunning, Chris Elbers, Antonie de Kemp and Phil Compernelle, “A contribution to a methodology for the Evaluation of Budget Support”, of December 2008 and owes a lot to the authors of this work. For the proposed impact evaluation of budget support, both papers borrow from Chris Elbers, Jan Willem Gunning and Kobus de Hoop (2008), “Assessing Sector-Wide Programs with Statistical Impact Evaluation: A Methodological Proposal”.

The present paper has tried to ‘simplify’ the previous paper of Gunning et al. The paper also borrows from draft guidelines for NONIE (the Network of Networks on Impact Evaluation) and especially from the final draft written by Frans Leeuw and Jos Vaessen. The paper benefited furthermore substantially from the valuable comments of the core group for the “Evaluation of Budget Support Operations at Country Level”, gathered during the meeting on December 9th, 2008, as well as from the written comments from Catherine Pravin and Chrysostomos Tsinas (European Commission), Remy Beaulieu (CIDA), Andrew Lawson, Martin van der Linde and Enzo Caputo (authors of the “Issue Paper”) and from Geske Dijkstra (Erasmus University Rotterdam). Nevertheless, the responsibility for the content of the current paper rests exclusively with the current editors.

Phil Compernelle
Antonie de Kemp
The Hague
April 2009

Contents

Tools for "Step 2":.....	1
The evaluation of the impact of government strategies	1
1 Introduction.....	3
2 Development of a toolbox for ‘Step 2’	4
3 Methodology for evaluating impact of government policy at sector level	5
3.1 Public Expenditure Tracking and Quality Surveys.....	6
3.2 Statistical impact evaluation (SIE).....	7
3.3 Beneficiary Incidence Analysis	9
3.4 Data.....	13
4 Methodology for evaluating impact of government policy supported by GBS.....	14
4.1 Introduction.....	15
4.2 Macroeconomic impact assessments	16
4.3 Cross-country comparisons.....	17
5 Qualitative approaches.....	18
6 Constraints and requirements.....	20
7 Conclusion	20
References.....	23
Annex 1. Statistical Impact Evaluation.....	26
Annex 2. Examples of Statistical Impact Evaluations	30

1 Introduction

During the past decade many donor countries have shifted from financing specific development projects to directly contributing to the government's budget without explicitly earmarking these contributions for specific activities. Such budget support can aim at a specific sector (sector budget support – SBS) or at the overall national development strategy (general budget support – GBS).¹ Parallel to this development, interest from the public and politicians has grown in assessing the effectiveness of foreign assistance.

The Joint Evaluation Unit of the European Commission for External Relations assigned a team of consultants (including DRN, ODI and ECORYS) to develop a comprehensive methodology for the evaluation of budget support operations at country level (May 2008). The extensive Issue Paper gave a brief overview of recent evaluations and methodologies and presented a comprehensive evaluation framework and a “three step approach” for the evaluation of GBS (p. 13-17):

- The *first step* focuses on the relevance of the inputs provided, the direct outputs of these inputs, and the quality and adequacy of the changes supported in the government systems (induced outputs: e.g. strengthened PFM).
- The *second step* involves a broad assessment of the GBS/SBS-related outcomes and impacts of the government strategies. This provides exhaustive information on the achievement of developments results and on the related determinant factors.
- The *third step* combines and compares the results of the first and second steps. It gives an assessment of the contribution of GBS/SBS to the factors that have had a key role in determining the success or failure of the government strategy.

As a complement to the Issue Paper, this paper proposes some evaluation tools for assessing Step 2; more specifically: evaluating the outcomes and impact of government policies (or induced policies in the Issue Paper), financed (partly) by budget support. Obviously, this is only one component of a comprehensive evaluation of budget support. For the other components, e.g. the impact of budget support on the institutional framework for public spending (PFM) or the role of policy dialogue, methodologies exist and many have been tested in the previous OECD-DAC Joint Evaluation of GBS, as described more elaborately in the Issue Paper.

Section 2 discusses in more detail the proposal to evaluate the outcomes and impact of government policy supported by budget support, using quantitative and qualitative evaluation tools (or Step 2 from the Issue Paper). Section 3 proposes a methodology for evaluating the impact of budget support at sector level. Section 3.1 sketches the use of expenditure tracking and quality surveys as a starting point. Section 3.2 explains the

¹ For definitions, see Issue Paper p.1-3

methodology for statistical impact evaluation of SBS. Section 3.3 addresses the question whether policies, or impact, are pro-poor, which can be studied through ‘benefit incidence analysis’ and section 3.4 pays attention to data requirements. Section 4 proposes a methodology for the broader evaluation of general budget support, moving beyond sector impact and adding the macro-economic and poverty impact. The mentioned techniques are mainly quantitative. Section 5 therefore introduces a brief discussion on qualitative approaches and their relation with quantitative analysis. Section 6 discusses several constraints and requirements for impact evaluations of budget support and section 7 concludes.

2 Development of a toolbox for ‘Step 2’

In the issue paper, ‘Step 2’ focuses on GBS/SBS-related outcomes and impacts of the government strategies.² For an evaluation of total ODA, Bigsten, Gunning and Tarp (2006) proposed to combine a qualitative approach with quantitative impact evaluation. The proposed method is that of country case studies with two stages. The first is an analysis of policy choice and implementation (similar to Step 1 of the Issue Paper) and the second stage involves a broad-based application of statistical impact evaluation at sector-level. The identification of sectors could be based for instance on the Poverty Reduction Strategy Paper (PRSP) or other national development strategies and associated result management frameworks (e.g. Policy Assessment Matrices).³ Non traditional sectors may also be part of the PAF (Performance Assessment Framework) reflecting the expected results of this strategy.

In practice, an impact evaluation of one or two sectors may already be demanding. Therefore an approach *might* be the combination of new impact evaluations for one or two sectors combined with an assessment of the effectiveness of other sectors, based on existing impact studies (at the sector level). With an increasing number of impact evaluations this approach may become more viable. The development of a (joint) research agenda for sector studies may be a good strategy for evaluating the effectiveness of budget support, based on a meta-evaluation of these studies.

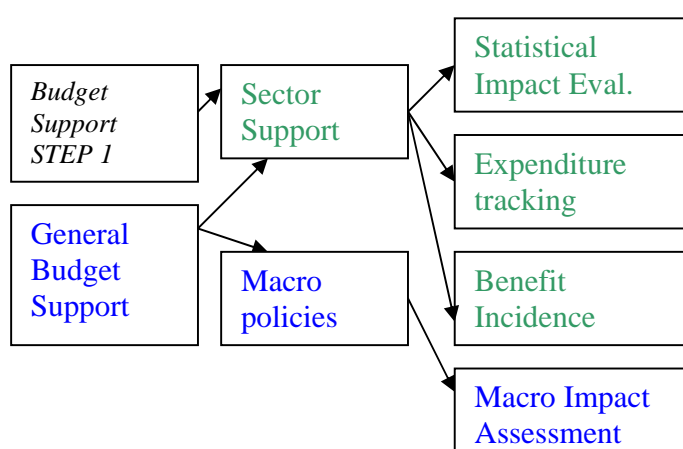
At the same time, it must be acknowledged that an assessment of the impact of GBS is more than a portfolio of SBS studies. There are policy elements for which the impact

² A specific phenomenon to take into account is fungibility. As long as the donor contribution to a sector’s budget remains limited, it is unlikely that much sector budget support ‘spills over’ to other sectors. Thus it is reasonable to link sector budget support to policy in that sector. However, even at the highest level of general budget support one should anticipate fungibility, in particular between foreign assistance and other forms of government revenue, such as taxes, and various elements of government expenditure, such as subsidies of food or fuel, that are determined at the central government level.

³ Impact evaluation is of course not limited to the social sectors; see the example of rural electrification in Annex 2.

occurs for the country as a whole, for example, in the case for trade policy, regulation and competition policy. Fortunately, different methods have been developed for the analysis of these types of policy.⁴ These range from relatively simple qualitative methods (e.g. assessing the effect of changes in agricultural prices on poverty by using survey evidence on whether poor groups are net buyers or sellers) to highly complex quantitative techniques (e.g. computable general equilibrium simulations to assess the distribution impact of trade policy).

Figure 1: *Evaluating Budget Support STEP 2: Impact of government policy*



As part of Step 2 of the Issue Paper, this paper starts with a methodology for evaluating the impact of sector budget support. The main approach is statistical impact evaluation, complemented with expenditure tracking and benefit incidence. Next, this approach is expanded to evaluate the impact of general budget support (again only for Step 2 of the Issue Paper) by adding to the approach for sector budget support, a proposal for evaluating the impact on macro-economic variables.

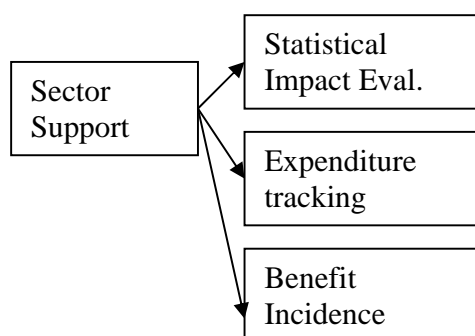
3 Methodology for evaluating impact of government policy at sector level

Figure 2 illustrates the proposed approach to evaluate sector budget support. The main component is statistical impact evaluation, which is currently the standard for evaluating the impact of activities (3.3). However, as a starting point Public Expenditure Tracking and Quality Surveys can be very informative (3.1). Benefit incidence analysis further supplements the evaluation of sector budget support, by addressing the question of reaching the poor (3.2). This list of methodologies is certainly not exhaustive. Moreover,

⁴ See e.g. Winters et al. (2004).

though this paper does not cover qualitative methods, the proposed methodologies will certainly have to be accompanied by qualitative analysis.

Figure 2: *Evaluating Sector Support STEP 2: Impact of government policy*



3.1 Public Expenditure Tracking and Quality Surveys

An impact evaluation focuses on the relation between output and outcome or on the relation between output and impact. Nevertheless, one may need to analyse the outputs first. For example, textbooks may be an effective instrument for improving learning, but does the centrally provided money actually reach the schools so that they can buy textbooks? In the evaluation framework proposed in the Issue Paper, this question lies on the intersection of Step 1 and Step 2, and can be seen as a starting point for Step 2.

A *Public Expenditure Tracking Survey* (PETS) provides an answer to this question. This survey traces the “ways in which public expenditures become public goods” by examining service facilities. It can provide information on the actual allocation of resources at different government levels and locate the leakage of funds (Dehn, Reinikka and Svensson, 2003). A related instrument is the *Quantitative Service Delivery Survey* (QSDS). This survey of facilities and service providers (e.g. health clinics, schools) analyses the efficacy of public spending and the incentives for frontline service providers to deliver.⁵ Both tools use a multiangular data collection strategy, combining data from different sources as a means of cross-validating the information from each source. This might be rather time consuming, though for many countries the information is already available for use in a budget support evaluation.

⁵ For more information, see the site of World Bank Human Development and Public Services Research <http://go.worldbank.org/1KIMS4I3K0>

Box 1: Expenditure tracking in Uganda

A tracking survey for Uganda showed that only 13 percent of non-wage recurrent expenditures for primary education actually reached primary schools in 1991-95 (Reinikka and Svensson, 2004). Moreover, larger schools and schools with wealthier parents received a larger share of the intended funds (per student) than smaller schools and schools with poor parents. The variation in grants received across schools was determined more by the political factors than by efficiency and equity considerations.

3.2 Statistical impact evaluation (SIE)

Statistical impact analysis is rapidly becoming the standard approach when the evaluation concerns government interventions that supported a range of activities. The evaluation of PROGRESA, a subsidy programme with educational grants for the poorest families in Mexico, is one of the most famous examples (see for instance Skoufias, 2005). Besley and Burgess (2000) used a regression-based approach to evaluate land reform in India, the Independent Evaluation Group (IEG) of the World Bank did so for educational reforms in Ghana (OED, 2004) and for child health in Bangladesh (OED, 2005). Other evaluation departments (especially IADB, ADB and AFD) are using these statistical techniques for their evaluations. In the Netherlands, the evaluation department of the Ministry of Foreign Affairs is working on evaluations of education and water and sanitation policies (e.g. IOB, 2007 and 2008). A perceived lack of 'rigorous impact evaluations' resulted in the creation of "3IE" and the international Network of Networks on Impact Evaluations (NONIE).

Most statistical impact evaluations, carried out in the past, were evaluations of projects. The evaluation of a whole sector poses new challenges to impact evaluations. At the project level, one may compare the results of a project with the development of a control group. At the sector level, such a counterfactual *appears* to be missing. Nevertheless, in many cases there is a way out. While sector policies normally have the same objectives for the whole country, the implementation of these policies over the country is phased and uneven. Government interventions are rarely applied in the same way in all areas of the country at the same time. Hospitals and schools are being built, roads constructed and electrification projects implemented, but not everywhere at the same time. Some schools may receive very few textbooks in a given year, others many and still other schools none at all. Similarly, a nation-wide teacher training or vaccination programme may initially affect only a few schools or villages, but eventually virtually all. At the sector level interventions are therefore *heterogeneous*: they differ across space and time. Elbers, Gunning and De Hoop (2008) propose to use this heterogeneity in the implementation of policies for a statistical impact evaluation of sector budget support (SBS) (see also IOB, 2008 for examples). The analysis assesses outcomes and impact of sector policies at the

level of (potential) beneficiaries of government service provision in a sector (households, school children, firms, etc.). Foreign assistance can claim to have a share in the impact to the extent that it has helped financing sector policies.

Box 2. Data sources for statistical impact evaluation

Some statistical impact evaluations have been based on primary data collection. For example, for the statistical impact evaluation of the education sector in Ghana, an additional education module was added to the Ghana Living Standards Survey. This module resurveyed 85 communities and their schools, which had been covered in the GLSS of 1988/89. In the interests of comparability, the same questions were kept, although additional ones were added pertaining to school management, as were two whole new questionnaires. The study thus had a unique data set (OED 2004).

The statistical impact evaluations of the education sector in Uganda and Zambia by IOB relied on the use of secondary (administrative data), though they included some field work as well (IOB, 2008a,b). The main sources are the annual school census data, national assessment tests, examination data, Demographic and Health Surveys (DHS), especially the EdData Surveys, the Population and Housing Census, SACMEQ II data. The Uganda evaluation also conducted a specific survey on teacher absenteeism. Many other statistical impact evaluations use existing data, as described in more detail in the annex.

The development of an analytical model is the first step in an evaluation to guide the data collection and analysis. This model can be based on existing frameworks used in country (e.g. results management frameworks, policy assessment matrices), but it is important to keep an open mind. There is a risk that the researcher focuses on the policy variables and intended effects, ignoring other determinants, selection bias, reverse causalities and unintended effects.⁶ Literature reviews, expert interviews and discussions and focus group discussions can be used both before survey data are collected (to ensure that relevant questions are included, notably about the unintended effects of policies) and after preliminary results from a quantitative analysis are available (as a reality check on the results).

Limitations

Statistical impact evaluation is not appropriate for policies that are the same everywhere in the country. This is normally the case for legislation and economic policy (for example a national competition policy). In that case one cannot assess its impact by comparing a

⁶ See e.g. World Bank (2000).

‘treatment’ and a ‘control’ group. More generally: there is no variance that can be exploited to estimate impact. It is important to stress that this is not a characteristic of statistical or econometric approaches. A qualitative analysis faces the same fundamental problem: a rigorous counterfactual cannot be constructed in this situation and an assessment of the ‘plausibility’ appears to be the best one can get. The existence of *unobserved* variables, leading to the risk of a selection bias is not a specific phenomenon of an econometric analysis. On the contrary: econometric techniques, such as the technique proposed in this toolbox, are in a better position to deal with this problem than many other techniques.⁷

Sometimes, the effects are obvious. In the case of Uganda, the enrolment to primary education doubled when the government introduced free primary education. But in the case of Zambia, the abolishment of school fees *apparently* had no effect. Does this mean that this measure failed to improve enrolments? An impact evaluation showed that enrolments did not increase immediately, because schools did not have the capacity to accommodate these children. Schools just refused to accept them. Gradually, with the improvement of the capacity in schools, the number of pupils increased. The example shows the importance of combining statistical and qualitative analyses. Qualitative techniques (case studies and interviews) help to formulate hypotheses about causes and effects; statistical techniques help to generalise the findings and to test assumptions about the attribution of instruments to the realisation of specific objectives (or unintended effects).

Annex 1 discusses the methodology for statistical impact evaluation more in detail.

3.3 Beneficiary Incidence Analysis

If budget support intends to contribute to the realisation of the objectives formulated in Poverty Reduction Strategy Papers (PRSPs), the evaluation should focus on the realisation of these objectives. Of specific interest would be the improvement of service delivery among the poorest groups and the reduction of poverty. For instance, one may want to analyse the distribution of expenditure on education or health among specific income groups.

Benefit incidence analysis provides a number of tools to analyse this (Bourguignon and Pereira da Silva, 2003). It assesses how the benefits of government spending are distributed across the population (Younger, 2003). Such an analysis can be done for those parts of government taxation and spending that can be readily associated with particular

⁷ This is also the conclusion of the (draft) guidance document of Leeuw and Vaessen (2009), written for NONIE (Network of Networks on Impact Evaluation).

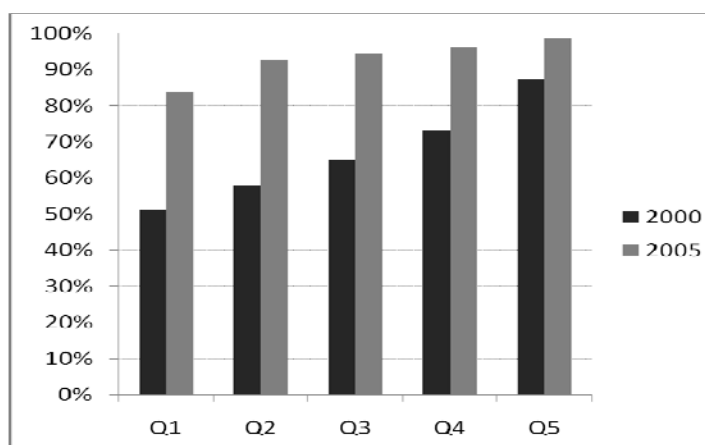
(groups of) individuals. Demery (2003) gives a description of the four steps, involved in a benefit incidence analysis:

1. the estimation of unit subsidies (the cost of providing the service divided by the number of beneficiaries);
2. the identification of users;
3. the aggregation of users into groups;
4. accounting for household spending.

As such, the cost of provision is used as a proxy for benefits received. This and some pitfalls, as well as solutions, have been discussed extensively by Demery (2003) and Younger (2003).

A simpler way to describe the distribution of government spending across groups of individuals is to give a presentation by wealth percentile or quintile. The information for such presentations usually comes from standard household surveys like Living Standards and Monitoring Surveys (LSMS) or Demographic and Health surveys (DHS). These surveys include information about the use of public services as well as information about household characteristics (including income).⁸ Empirical studies indicate, for instance, that the poorest quintiles tend to profit from primary schooling subsidies, while the richest quintiles profit from tertiary schooling subsidies (Demery 2003, p. 45). An evaluation of health spending in Ghana showed that the poorest quintile makes little use of publicly provided health care facilities in comparison with the better off (ibid, p. 53). Figure 3 below gives an example of education in Zambia. Clearly, over the period considered education spending became more pro-poor.

Figure 3: *Imputed Public Education Spending in Zambia by Income Quintile*

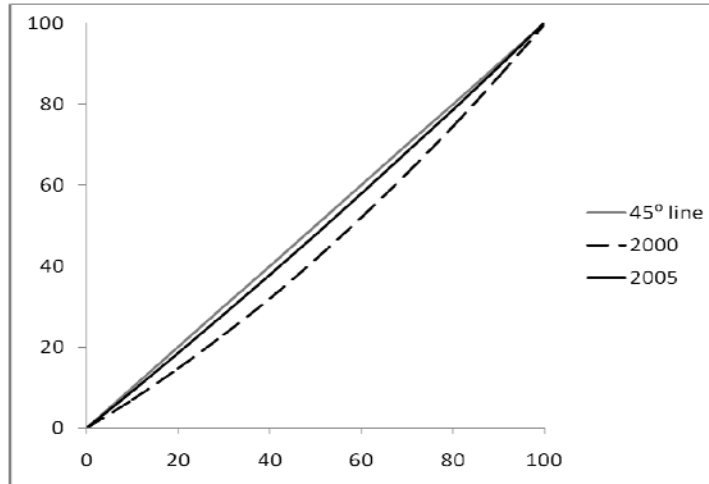


Source: IOB (2008b).

⁸ Occasionally, one might be in the position to add questions to such surveys that allow one to infer the direct distributional characteristics of a policy.

A central tool in incidence analysis is the ‘concentration curve’ (see Figure 4, again for Zambia in 2000-2005). For instance, an enrolment concentration curve shows the share in total enrolment for the poorest 10%, 20%, 30%, etc. of the population.

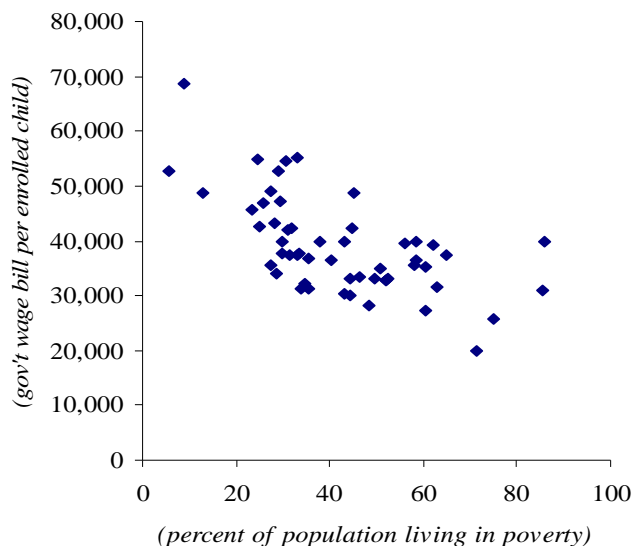
Figure 4: *Concentration Curve of Enrolment in Zambia, 2000-2005*



Computed from IOB (2008b).

A benefit incidence analysis can also focus on regional groups. For countries with a national strategy focused on poverty reduction, the improvement of service delivery and the reduction of poverty in the poorest regions would be of specific interest. For such an evaluation, researchers may use existing *poverty maps*. These maps combine information from household surveys with census data. Figure 5 below gives an example of such an analysis for primary education in Uganda, combining administrative data (at the district level) with poverty maps. It shows that education expenditure is regressive, with more money going to the richest districts.

Figure 5: *Government Education Wage Bill by District Poverty Level, Uganda*



Source: Winkler and Sondergaard (2007)

Limitations

The examples in Figures 4 and 5 illustrate the importance of a *dynamic* analysis to assess the effects of government policies. Often, most recent or future spending is benefiting a different group than previous spending. For instance, spending on schooling can be regressive on average, while *marginal* spending on schooling is actually pro-poor (Lanjouw and Ravallion, 1999, Younger, 2003).

A second warning is in place. While benefit incidence analysis may give information about the distribution of public services among groups of people or may register changes in this distribution over time, it fails to provide a well-defined counterfactual. For instance, in the Zambia example, a large part of the increase in enrolments in the poorest quintiles could be attributed to the growth of community schools. These schools are, however, not financed by the government. Benefit incidence analysis is therefore no substitute for a proper impact evaluation.⁹

⁹ In rare cases it may be possible to assess differential impact for different income strata, in particular if data for the evaluation have been collected as part of a nationally representative household survey (see also Younger, 2003). In such cases the incidence analysis of impact would provide a more complete answer to the question whether a particular policy (or the ensemble of sector policies) is pro-poor.

3.4 Data

Whatever the method and techniques used and whatever the source, the dataset should be representative for interventions and outcomes at the level of the ultimate beneficiaries, typically individuals or households. For example, in the case of education one may draw a sample of schools and make the probability that a school is included in the sample proportional to the population size of a school's catchment area. This guarantees that the interventions from which the schools in the sample benefited are representative (from the perspective of the population) for the interventions in the sector.

To get a representative dataset, evaluations of sector budget support may use specific surveys, but it is important to keep in mind that more and more administrative and census data become available, while standard (household) surveys may also be used for these evaluations.

Data sources at different levels

	Administrative data	Census data	Standard surveys	Specific surveys
Input	X			
Output	X			X
Outcome	X		X	X
Impact	X		X	X
Non-policy variables		X	X	X
Beneficiary incidence analysis	X	X	X	X

A PETS may need an additional survey, but may partly rely on administrative data. Tracking surveys have been undertaken in many countries and may be available. For a beneficiary incidence analysis administrative data, census data and standard household surveys may be used. An impact evaluation may need data at output, outcome and impact level and rely on a combination of administrative data, census data, standard surveys and specific surveys. For the evaluation sample, three types of data must be collected:

- most relevant impact variables, including intended as well as unintended effects;
- interventions (at the level of beneficiaries) in the sector;
- all observable non-policy determinants of impact.

For the *output variables*, intervention histories have to be constructed: measures of how and when the sample was affected by government interventions. At this level, it is probably possible to use administrative data.

Outcomes are obviously determined by many factors other than the policies recorded in the intervention history. In principle they should all be included in the regression.¹⁰ Sometimes it is possible to use secondary data, i.e. administrative data or existing surveys. School census data, for instance, include information about pupils and examination boards may have information of test- and examination results. With triangulation, linking the school census data to other databases (as assessment and household surveys) it is possible to include other outcome variables and to check the reliability of the census data.

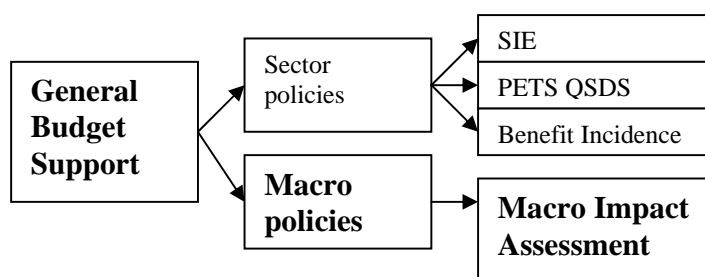
The *impact variables* must include measures for the *intended* effects of the sector policies and preferably also measures *unintended* effects. Demographic and Health Surveys contain a lot of information on the use of public services at the household level. Nevertheless, in many cases additional surveys might be required.

4 Methodology for evaluating impact of government policy supported by GBS

As far as GBS is expected to have an impact on sector policies (and especially the social sectors), it is appropriate to use the same evaluation tools as for SBS. Nevertheless, the goals and impact of GBS are wider and go beyond the sector level. This section focuses on the impact of macro policies.

Figure 6 sketches an approach for evaluating the impact of government policies, supported by general budget support. It is based on developing different scenarios to identify plausible impacts of macro policies through quantitative and qualitative methods. Again, it is important to note that this suggested approach covers only part of a comprehensive evaluation (which would cover not only Step 2 but also Step 1 and 3).

Figure 6: *Evaluating General Budget Support STEP 2: Impact of government policy*



¹⁰ Even if omitted variables are strongly exogenous it may still be useful to include them to get more precise estimates.

4.1 Introduction

An evaluation of the effectiveness of General Budget Support should focus on the effectiveness of policies that are supported by GBS. The broader aid environment is important as well (e.g. other aid delivery mechanisms, country ownership and governance, donor harmonisation). Moreover, budget support is not a fixed variable but will change over time (e.g. tranches), whereby timing of commitments, disbursement and actual expenditure also matters. Any evaluation will therefore have to start with analysing changes in government's budget and financing to trace the flow of funds. This includes addressing saving and spending of aid, shifts in sector allocations and changes in other sources of government revenue following budget support. Step 1 of the common framework proposed in the Issue Paper deals with these subjects, which are considered "induced outputs" of GBS/SBS.

With regard to Step 2, the outcome and impact of budget support related strategies, one can distinguish two main (overlapping and interacting) areas of macroeconomic impact of budget support:

1. response by the general economy to the way aid is received and absorbed (e.g. impact on balance of payments, factor prices, labour market, interest rates, allocation of resources private and public sector);
2. impact on sustainable economic growth and poverty reduction of aid-financed government expenditure and investments:
 - i. direct impact through economic sectors, e.g. infrastructure, private sector development programs and distributional policies;
 - ii. indirect (longer term) impact through social sectors.

It is important to note that, on the one hand, the sector or distributional impact of budget support will affect macro variables through the behavioural responses at micro-level, e.g. effect of education on labour supply and economic growth (in the longer run). On the other hand, changes in macro variables will also affect distribution and outcomes in the social sectors. The challenge is to capture these macro-micro linkages¹¹.

Options

A problem when evaluating the macroeconomic impact is the absence of a clear counterfactual. Macroeconomic policy is national and applies to the whole country (or region as in e.g. India). There is, nevertheless, an enormous amount of literature on the

¹¹ An evaluation of budget support will also need to take into account the time lags required for impact to be realised. For example, budget support will have a direct impact on the exchange rate and monetary policy, but a much slower impact on the supply-side of the economy.

analysis of macroeconomic impact (see Bourguignon, 2003, and Winters *et al.*, 2004). In this section we discuss two methods that include the construction of a counterfactual:

1. estimating the effects through the *simulation* of a situation where aid or policy changes are absent;
2. *cross country comparisons*: using a comparison with other countries

4.2 Macroeconomic impact assessments

A first way to construct a counterfactual is through the analysis of different scenarios, possibly making use of periods with drops or changes in aid to identify plausible impacts through quantitative and qualitative methods.

White and Dijkstra (2003) give examples of the evaluation of aid using such counterfactual analysis. The authors trace for nine country case studies how program aid has affected macro-economic aggregates (like imports and government spending) and (through these indicators) economic growth. The analysis involves qualitative and quantitative methods and combines a description of the policy dialogue and government policies with an evaluation of the impact of aid (donor funds and dialogue). The authors analyse marginal impact in three areas: the external account, the internal account and government finance. The counterfactual analysis is carried out for the balance of payments (simulating a situation without aid, debt forgiveness and/or debt service).

The analysis of different scenarios (with / without or with less aid) is also possible using more complicated macroeconomic simulation models. These models, often Computable General Equilibrium (CGE) models, normally focus on the (future) effects of policy or exogenous changes and are widely used for macroeconomic analysis and planning purposes. Bourguignon and Da Silva (2003, 2008) give a comprehensive overview of available models used to forecast the impact of macro-level policies on poverty and income distribution. Such models have been used to analyse the effect of policy changes, such as adjustment policies, on the income distribution and poverty levels in developing countries.

However, these models could also be used for *ex post* evaluation in country case studies comparing the outcome of budget support (model calibrated to fit most recent year with budget support) with an alternative scenario of no-budget support (less aid or aid delivered differently). It may even be possible to link some aspects of micro-impact (e.g. education and labour market) back to the macro economy. Micro/macro models that include representative household groups (RHs) make it possible to analyse simultaneously changes both in the structure of the economy and in the distribution of income.

The World Bank has constructed a dynamic country-level model, MAMS (Maquette for MDG Simulations) to address questions about the effects of different government and foreign aid policies on social and economic performance (Lofgren 2007). The performance indicators cover poverty and other MDG targets. For many countries, this or similar models have already been developed, often with assistance of international organisations such as IMF, UNDP and the World Bank.¹² Where they exist, they might be a valuable tool for evaluating the impact of budget support on economic growth and poverty reduction. The results of the evaluation of Step 1 would provide useful information to inform the different possible scenarios about how the government would have allocated resources in the absence of the budget support. Moreover, these kinds of tools might also be informative for Step 3.

Limitations

For a correct interpretation, it is important to have a good understanding of these models. The scenarios have to be based on sound understanding of the country context (preferable the result of a workshop with representatives from government and donors). Moreover, the estimated effects depend critically on the assumptions used in the equations, so these need to be assessed carefully. The parameters in the model are sometimes based on a 'best guess' or are 'calibrated'. Clearly, the more use is made of estimated parameters the more convincing the results of such simulations will be.

4.3 Cross-country comparisons

A second option would be to try to compare countries. There is a large body of literature on cross country studies. The empirical cross-country studies rely heavily on regression analysis, using the variation in growth rates and explanatory variables to analyse the causes of differences in growth rates (see for instance Barro and Sala-i-Martin, 2004). The same technique for cross country case studies has also been applied to an analysis of the effectiveness of aid.¹³ An important element of the technique is the (implicit) creation of a counterfactual and the inclusion of many relevant determinants of economic growth, apart from government policies and the effectiveness of aid.

¹² For countries, see <http://go.worldbank.org/OPBK6FI290>. For general information on MAMS, see <http://go.worldbank.org/XSQT186EN0>

¹³ See for instance Burnside and Dollar (2000 and 2004), Collier and Dollar (2002 and 2004) Hansen and Tarp (2001), Clemens, Radelet and Bhanvnani (2004) Rajan, Raghuram Subramanian (2005).

Limitations

When using this kind of approach, it is important to be aware of the challenges. As a result of a small (sample) size, outliers and influential data points have a large effect on the estimates (see e.g. Roodman, 2007 and 2008). The inclusion of aid variables also needs specific attention. A large part of aid does not aim at having an impact on economic growth or poverty reduction in the short run (Clemens *et al.*, 2004).

A more qualitative approach, comparing two or three countries, would also be possible. Nevertheless, one should be aware that attribution problems are even much larger when using such an approach. One cannot simply compare countries receiving budget support with countries receiving project aid or no aid at all. Reverse causalities play an important role. Donors and partner countries decided to move from project aid to sector or budget support, because they felt the conditions were right. The more qualitative analysis of Step 1 will be extremely important to analyse the policy dialogue and contextual differences.

In general, an evaluation of the macro impact of GBS will therefore have to draw on multiple sources, both qualitative and quantitative. A promising approach appears to be the combination of cross-country regressions with in depth country case-studies. Ndulu *et al.* (2008) give an excellent example of such an approach for studying economic development in Africa. The main report analyses political and economic developments in Africa combining (cross country) regressions with a detailed qualitative analysis. The second volume incorporates 26 country case studies, used to validate and interpret the results of the cross country regressions. This approach might be useful and interesting as part of the synthesis of the findings of the SBS/GBS evaluation. Of course, this approach presupposes a larger number of country case studies.

5 Qualitative approaches

Qualitative methods are normally process oriented and stress the role of actors. Examples are logical framework approaches or participatory approaches. Qualitative methods are flexible, often by using an eclectic approach and may contextualise empirical findings. Qualitative research is important for an understanding of causal relations and may contribute to an assessment of the *plausibility* of the effectiveness and impact.

The (draft) NONIE guidance document (Leeuw and Vaessen, 2009) gives an excellent overview of quantitative and qualitative techniques. The main qualitative approaches are, apart from (qualitative) case studies and (open and semi-structured) interviews:

- *participatory approaches*: these are built on the (normative) principle that stakeholders should be involved in some or all stages of the evaluation. Advantages

of a participatory approach may be more ownership and a better understanding of processes of change and the way in which interventions affect people;

- *causal contribution analysis*: “contribution analysis relies upon chains of logical arguments that are verified through a careful analysis. Rigor in causal contribution analysis involves systematically identifying and investigating alternative explanations for observed impacts. This includes being able to rule out implementation failure as an explanation of lack of results, and developing testable hypotheses and predictions to identify the conditions under which interventions contribute to specific impacts” (Leeuw and Vaessen 2009, p. 25-26). The authors show how the causal story is inferred from the following evidence:
 1. there is a reasoned theory of change for the intervention: it makes sense, it is plausible, and is agreed by key players;
 2. the activities of the intervention were implemented;
 3. the theory of change—or key elements thereof— is verified by evidence: the chain of expected results occurred;
 4. other influencing factors have been assessed and either shown not to have made a significant contribution or their relative role in contributing to the desired result has been recognized.

Limitations

Qualitative approaches normally focus on small groups and this makes it difficult to generalise the findings of these studies. Moreover, it may be difficult to disentangle the role of the various factors that contribute to the overall effect. Limitations of participatory approaches are limitations to the validity of information based on stakeholder perceptions (only) and risks of strategic responses and manipulation by stakeholders. Causal contribution analysis has the same limitations: “Despite the potential strength of the causal argumentation on the links between the intervention and impact, and despite the possible availability of data on indicators, as well as data on contributing factors (etc.), there remains uncertainty about the *magnitude* of the impact as well as *the extent* to which the changes in impact variables are really due to the intervention or due to other influential variables.” (Leeuw and Vaessen 2008, p. 26).

Quantitative methods are better suited for the measurement of effects and impact and the generalisation of findings. With statistical or econometric techniques it is possible to isolate the effect of interventions in a complex environment, where many factors – beyond the intervention itself – have an impact. These methods have their price as well: they can be applied only when the samples are large enough.

Quantitative methods do not replace qualitative methods. On the contrary. Good research combines quantitative and qualitative methods and techniques. Good evaluations are almost invariably mixed method evaluations and include qualitative and quantitative

information.¹⁴ A good evaluation starts with a theory-based design. Without such a design one gets *black box evaluations*: they are able to ‘measure’ impact, but give no clue as to why the intervention is or is not giving the expected impacts. Moreover, without a proper theoretical design, there is a risk of *data-mining* with an arbitrary looking for correlations. Such an approach has nothing to do with evaluation and is inconsistent with the underlying statistical probability theory.

6 Constraints and requirements

1. Like any evaluation, the evaluation team must consist of members with the technical competence to assess and solve the specific methodological problems, as well as members with specific knowledge of the sector (see also Leeuw and Vaessen, 2009).
2. The evaluation design must be clear on the sources of data and realistic about how long it will take to collect and analyze data. The process of analysing the quality, the processing and linking of *secondary data* may take up to three months. Moreover, it is important to include time (three months) for getting the data, even though they should be already available somewhere.
3. It is important to analyse the availability and quality of data. Note, however, that this is not a specific characteristic of quantitative research. Surveys are not reliable by definition. The same holds for more qualitative information, obtained for instance through interviews. Triangulation is an important instrument for the analysis of the quality of information, whether quantitative or qualitative.
4. The *data analysis* may take approximately three to four months. The researchers will encounter unexpected problems that will have to be solved during that analysis. This may even require additional data.

7 Conclusion

New aid modalities have important implications for the evaluation of the effectiveness of bilateral support. There is quite a lot of experience with doing impact evaluations of specific projects. There is less experience with conducting evaluations in the context of budget support. As a result, the evidence on the impact of budget support is rather scanty. At the same time, evaluation departments and units face demands from many sides – including national parliaments and European institutions – to give more insight into the effects and impact of general budget support. This requires further development of a methodology for the measurement of the results of budget support.

¹⁴ For further discussion see Michael Bamberger 2006.

This paper seeks to contribute to the approach for the evaluation of the impact of budget support as described in the Issue Paper. The paper focuses on the effects and impact of government policies, financed (partly) by budget support. This corresponds with Step 2 of the proposed common framework in the Issue Paper. Complementing the Issue Paper, this paper examines the strengths and limitations of quantitative methodologies for evaluating the impact of government strategies financed through sector or general budget support and proposes the use of these techniques whenever they are applicable. These methods enable the rigorous measurement and quantification of the effects of budget support where this is possible.

The proposed techniques exploit the heterogeneity at the sector level to construct a counterfactual. The evaluation of general budget support could include statistical impact evaluation of the main relevant sectors, chosen from the country's national strategy e.g. the Poverty Reduction Strategy. However, it is impossible to cover all sectors. A solution is the combination of existing impact evaluations at the sector level with one or two new (statistical) impact evaluations of other sectors. This suggests that the choice of sectors and countries for statistical evaluation studies should be guided by what is already available so that the sector coverage is as broad as the available budget allows. Moreover, the development of a (joint) research agenda for sector studies may be a good strategy to increase coverage of relevant sectors, based on a meta-evaluation of these studies.

Obviously, this heterogeneity is absent at the macro level of a country. Nevertheless, there appear to be two ways to construct a counterfactual: comparing different countries or comparing with a hypothetical alternative situation (for instance without budget support). Both methods may be combined with qualitative in-depth analysis in country case studies. For the specific purpose of evaluating budget support at the country level, we advise simple counterfactual analysis (like White and Dijkstra) or the use of existing models such as the World Bank's MAMS.

In some cases, constructing a convincing counterfactual might not be possible, in which case an assessment of the 'plausibility' of effects, based on qualitative *and* quantitative information, appears to be the best one can get. These approaches are not an alternative when more rigorous techniques are available: "... quantitative methods are usually preferable and should be pursued when possible, and qualitative techniques should be used to evaluate the important issues for which quantification is not feasible or practical" (Leeuw and Vaessen 2009, p. 42).

Finally, a warning seems to be in place. Evaluation of budget support bears the risk of *a lack of focus*. It is impossible to analyse the overall effects of donor policies and the overall effects and impact of government policies at the macro and the micro level at the same time in a rigorous way. The breadth of such an evaluation will result in a lack of

focus and therefore the results will be at the expense of one of the depth. For such an overall evaluation, the use of a meta-evaluation, appears to be a more promising approach.

References

- Barro, R. J. and X. Sala-i-Martin (2004), *Economic Growth* (second edition), Cambridge: The MIT Press.
- Besley, T. and R. Burgess (2000), ‘Land Reform, Poverty Reduction, and Growth: Evidence from India’, *Quarterly Journal of Economics*, vol. 115, pp. 389-430.
- Bigsten, A., J.W. Gunning and F. Tarp (2006), *The effectiveness of total ODA: an evaluation proposal*, Göteborg / Amsterdam / Copenhagen.
- Bourguignon, F. and L.A. Pereira da Silva (eds.) (2003), *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation techniques and tools*, Oxford: Oxford University Press.
- Bourguignon, F., M. Bussolo and L.A. Pereira da Silva (eds.) (2008), *The Impact of Economic Policies on Poverty and Income Distribution: Macro-Micro Evaluation Techniques and Tools*, Oxford: Oxford University Press.
- Caputo, E., A. Lawson and M. van der Linde (2008), *Methodology for Evaluations of Budget Support Operations at Country Level, Issue Paper*, DRN-ADE-EC-NCG-ECORYS.
- Clemens, M.A., S. Radelet and R. R. Bhavnani (2004), *Counting chickens when they hatch: The short-term effect of aid on growth*, Washington DC: Center for Global Development, working paper 44.
- Deaton, A. (1997), *The Analysis of Household Surveys: a Microeconomic Approach to Development Policy*, Baltimore: Johns Hopkins University Press.
- Demery, L. (2003), Analyzing the incidence of public spending, in: Bourguignon, F. and L.A. Pereira da Silva (eds.), *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation techniques and tools*, Oxford: Oxford University Press, pp. 41-68.
- Elbers, C., J.W. Gunning and K. de Hoop (2008), ‘Assessing Sector-Wide Programs with Statistical Impact Evaluation: A Methodological Proposal’, *World Development*, DOI:10.1016/j.worlddev.2008.01.002.
- Galdo, V. and B. Briceño (2005), *An impact evaluation of a potable water and sewerage expansion in Quito: Is water enough?* Washington DC.
- Gunning, JW, C. Elbers, A. de Kemp and P. Compernelle (2008), *A contribution to a methodology for the Evaluation of Budget Support*” (mimeo).
- IEG (2008a), *What works in water supply and sanitation? Lessons from impact evaluations*, Washington DC: World Bank.
- IEG (2008b), *The Welfare Impact of Rural Electrification: A Reassessment of the Costs and Benefits*, Washington DC: IEG Impact Evaluation Series, 2008.
- IOB (2007), *Water Supply and Sanitation Programmes: Shinyanga Region, Tanzania, 1990-2006*, IOB Impact Evaluation no. 305, The Hague: Ministry of Foreign Affairs, Policy and Operations Evaluation Department.

- IOB (2008a), *Primary education in Uganda*, Impact Evaluation no 311, The Hague: Ministry of Foreign Affairs, Policy and Operations Evaluation Department.
- IOB (2008b), *Primary education in Zambia*, Impact Evaluation no 312, The Hague: Ministry of Foreign Affairs, Policy and Operations Evaluation Department.
- IOB (2008c), *Support to Rural water Supply and Sanitation in Dhamar and Hodeidah Governates, Republic of Yemen*, IOB Impact Evaluation no. 315, The Hague: Ministry of Foreign Affairs, Policy and Operations Evaluation Department.
- Jalan, J. and M. Ravallion (2003), ‘Does piped water reduce diarrhoea for children in rural India?’ In: *Journal of Econometrics* 112(1), pp. 153-173.
- Joint Evaluation of General Budget Support 1994-2005 (2007), *Evaluation of General Budget Support – note on approach and methods*, London: DFID.
- Lanjouw, P. and M. Ravallion (1999), ‘Benefit Incidence, Public Spending Reforms, and the Timing of Program Capture’, *World Bank Economic Review*, vol. 13, pp. 752-773.
- Leeuw, F. and J. Vaessen, 2009, *Impact evaluations and development, NONIE Guidance on Impact Evaluation*, Maastricht (draft).
- Lofgren, H. (2007), Project note: *Development strategy analysis with MAMS*, Washington DC: World Bank.
- Ndulu, B., S. A. O’Connell, R. H. Bates, P. Collier and Ch. C. Soludo (eds.) (2008), *The political economy of economic growth in Africa, 1960-2000*, Cambridge: Cambridge University Press.
- OED (2004) *Books, buildings, and learning outcomes: an impact evaluation of World Bank Support to basic education in Ghana*, Washington DC: OED, The World Bank.
- OED (2005), *Maintaining Momentum to 2015? An Impact Evaluation of Interventions to Improve Maternal and Child Health and Nutrition in Bangladesh*, Washington DC: Operations Evaluation Department, World Bank.
- PREM (2006), *A guide to water and sanitation sector impact evaluation*, Washington DC: World Bank.
- Ravallion, M. (2001), ‘The Mystery of the Vanishing Benefits: an Introduction to Impact Evaluation’, *World Bank Economic Review*, vol. 15, pp. 115-140.
- Reinikka, R. and J. Svensson (2001), *Explaining Leakage of Public Funds*, Washington DC: World Bank Policy Research Working Paper No. 2709.
- Reinikka, R. and J. Svensson (2004), ‘Local capture: evidence from a central government transfer program in Uganda’, in: *The Quarterly Journal of Economics*, vol. 119 (2), pp. 678-704.
- Roodman, D. (2007), *The Anarchy of Numbers: Aid, Development, and Cross-country Empirics*, Washington DC: Center for Global Development, Working Paper 32.
- Roodman, D. (2008), *Through the looking glass, and what OLD found there: on growth, foreign aid, and reverse causality*, Washington DC: Center for Global Development, working Paper 137.
- Verbeek, M. (2000), *A Guide to Modern Econometrics*, Chichester: John Wiley.

- White, H. (2006), *Impact Evaluation: the Experience of the Independent Evaluation Group of the World Bank*, Washington, DC: World Bank.
- White, H. and G. Dijkstra (2003), *Programme aid and development, beyond conditionality*, London: Routledge.
- Winkler D. and L. Sondergaard (2007), *The efficiency of public education in Uganda*, Kampala: World Bank.
- Winters, L.A., N. McCulloch and A. McKay (2004), ‘Trade Liberalization and Poverty: The Evidence So Far’, *Journal of Economic Literature*, vol. 42, pp. 72-115.
- World Bank (2000), *The LogFrame Handbook*, Washington, DC: World Bank.
- Wooldridge, J. M. (2002), *Econometric analysis of cross section and panel data*, Cambridge: Massachusetts Institute of Technology.
- Younger, S.D. (2003), ‘Benefits on the Margin: Observations in Marginal Incidence’, in: *The world Bank Economic Review*, Vol. 17, no. 1, pp. 89-106.

Annex 1. Statistical Impact Evaluation

Evaluations of a programme must deal with attribution problems and selection bias. In general, regression techniques may be used to analyse the attribution of an intervention to a measured result. The problem is, however, that the results of a regression analysis may be biased in the case of *unobserved* characteristics, leading to selection effects. A simple example is the existence of (unobserved) differences between control group and intervention group. Irrigation projects, for instance, seem to have positive welfare impacts when the income of farmers in irrigated areas is compared with the income of other farmers. Nevertheless, selection effects are here lurking as well. Irrigated areas have normally a higher population density, a more developed population with better access to markets and a more fertile soil.

Double difference

The technique of double difference (or difference in difference) deals with these differences *as long as they are time invariant*. The technique measures differences between the two groups, before and after the intervention (therefore the name: double difference):¹⁵

	Intervention Group	Control Group	Difference across groups
Baseline	I_0	C_0	$I_0 - C_0$
Follow-up	I_1	C_1	$I_1 - C_1$
Difference across time	$I_1 - I_0$	$C_1 - C_0$	<i>Double-difference:</i> $(I_1 - C_1) - (I_0 - C_0) =$ $(I_1 - I_0) - (C_1 - C_0)$

Suppose there are two groups, an intervention group I and a control group C. The effect of an intervention may be estimated by measurement before (0) and after the intervention (1), for the intervention group (I) and for the control group (C). The effect of the intervention is:

$$(I_1 - I_0) - (C_1 - C_0) \text{ or } (I_1 - C_1) - (I_0 - C_0)$$

¹⁵ Adapted from: J.A. Maluccio and R. Flores (2005). See also White (2006).

Example

If enrolment rates at $t=0$ would be 80% (for the intervention group) and 75% for the control group and at $t=1$ these rates would be respectively 90% and 75%, then the effect of the intervention would be: $(90\% - 80\%) - (75\% - 70\%) = 5\%$.

This simple technique can only be used for binary situations: one can distinguish “treatment” and “control” groups and treatment is the same for all treated individuals. To take an example from the education sector, the intervention to be evaluated might be a conditional cash transfer program. The treatment group would then consist of the households receiving such transfers and a control group of households not receiving these grants.

This simple approach cannot be used to evaluate support for sector programmes or general budget support. For instance, an educational policy package contains many different types of interventions: construction of school buildings, provision of teaching materials, training of teachers, cash transfers to increase enrolment. These interventions affect the beneficiaries in many ways and in different degrees. In principle one could imagine doing a separate evaluation for each policy intervention and adding up the results of each to determine the impact of a policy package. However, results for individual interventions are bound to be affected by the presence and intensity of other policy interventions. In a regression this can be taken into account.

Regression analysis and the technique of differencing

Let impact variable Y_{it} depend on a vector of policy variables P_{it} , a vector of determinants not related to policy (“control variables”) X_{it} and a ‘disturbance’ term $\mu_i + \varepsilon_{it}$ explained below:

$$Y_{it} = a + bP_{it} + cX_{it} + \mu_i + \varepsilon_{it}. \quad (1)$$

Here i denotes the unit of the analysis (a school, a pupil, a household or a firm), and t the time of observation. A good measure for the impact of policy variables is the coefficient vector b , so the evaluation problem is reduced to estimating b . Typically, the coefficient vector b cannot be estimated by means of ordinary least squares (OLS) regression. The disturbance term $\mu_i + \varepsilon_{it}$, representing all variables omitted from the analysis, allows for a ‘fixed’ (i.e., constant over time) effect μ_i reflecting the possibility that units differ in

outcomes even if they do not differ in P or X . Such fixed effects are known to invalidate the results of simple regression techniques when they are correlated with intervention variables: they produce biased estimates of the coefficient vectors b and c .¹⁶ One way to deal with fixed effects is to ‘difference’ the regression equation:¹⁷

$$Y_{it+1} - Y_{it} = a + b(P_{it+1} - P_{it}) + c(X_{it+1} - X_{it}) + (\varepsilon_{it+1} - \varepsilon_{it}), \quad (2)$$

so that the fixed effect drops out of the equation.

Equation (2) is quite similar to the familiar ‘difference-in-differences’ estimator of more conventional policy evaluation (see above). The vector of impact coefficients b can now be estimated consistently if the changes in P and X are not correlated with the change in the disturbance term ε . An alternative sufficient condition for consistent estimation of b is that P reflects truly exogenous policy. Equations such as (2) can be estimated for all Y -variables, both for intended and unintended effects.

For example, the health status of the population may be affected by the extent to which the community offers suitable conditions for mosquito breeding. Differences between locations in health may therefore reflect differences in such conditions rather than in health policy interventions. However, since the conditions are likely to change only very slowly, if at all, the differencing procedure (regressing *changes* over time in health on *changes* in health policy) will produce an unbiased estimate of the impact of health policy (though admittedly one does lose information on levels). If the differences are not constant but can be described by a time trend, time should be included in the difference regression. This leaves the residual category of unobservable factors of influence, which are not constant and which cannot be described by a time trend. This category is problematic. Hence a serious effort must be made to ensure that all possible determinants are either observed or eliminated through differencing.¹⁸

In principle the impacts can be disaggregated by groups of recipients, notably to isolate the impact on the poor. This can be done by using the sample information on the distribution of interventions and combining these with the regression estimates of the impact of the interventions.

¹⁶ For a technical discussion of fixed effects, see e.g. Verbeek (2000, chapter 10) and Wooldridge (2002), chapter 10.

¹⁷ Besley and Burgess (2000) use a reduced form equation similar to equation (1). They have data for 30 years and are therefore able to estimate fixed effects at the level of the primary sampling unit so that there is no need for differencing. In sector evaluations time series are often quite short necessitating the differencing method we adopt in equation (2).

¹⁸ If the intervention is binary (as in conventional evaluation studies with distinct treatment and control groups) then the problem is often addressed by using matching techniques, e.g. propensity score matching.

Limitations

The technique of differencing depends very much on the quality of the data. An ordinary regression (like OLS) can deal with a certain measurement error, but differencing of the data reduces the 'signal noise ratio', or increases the measurement error over the signal. Second, the estimation becomes more sensitive to an incorrect treatment of time-lags. Moreover, when the main variables are "slow-moving", they will not generate much variation in differenced form when the time interval is not very large. As a result, the computed coefficients may become insignificant (Bates 2008, p. 273-274). Finally, the procedure stresses the time-series information within countries at the detriment of cross-sectional information and this results in a loss of precision (Barro 1998, p. 37 and 41 and Temple 1999, p. 132). In these cases, a better solution may be to try to find indicators (or proxies) for variables that may be expected to lead to a selection bias if not included in the equation.

Annex 2. Examples of Statistical Impact Evaluations

Impact evaluations of the education sector

Introduction

Education is one of the most evaluated sectors. Most evaluations are at the project level and focus on specific topics such as a specific teaching programme or teacher and pupil absenteeism. Nevertheless, with the shift towards sector support and the attention for the MDGs in education, impact evaluations focus increasingly on the whole sector or a subsector (like basic education). Examples are the IEG evaluation of education in Ghana (OED, 2004) and IOB evaluations of primary education in Uganda and in Zambia (IOB, 2008a and 2008b). In Ghana, the World Bank supported a range of activities in the basic education sub-sector, from rehabilitating school buildings to assisting in the formation of community-based school management committees. The IOB evaluations started from the premise that an analysis of the impact of sector support is an analysis of the impact of the policy *to which a specific donor or cooperating agency contributes*.

Data and methodology

The studies adopted a regression-based approach which analyzed the determinants of school attainment (years of schooling, progression, completion), specific problems (absence and repetition) and achievement (learning outcomes, i.e. test and examination scores).

The evaluations combined the econometric analysis with a description and analysis of the development of the education sector. In the IEG evaluation, for instance, the analysis of the political economy of education reform in Ghana was a vital piece of the story. The methodology thus adopted a theory-based approach to identify the channels through which a diverse range of interventions were having their impact and had much attention for the context. There are also large differences. One of these differences is the use of data. The IEG evaluation used the *Ghana Living Standards Survey* of 1988/89 as a baseline. This survey had an additional education module, which administered math and English tests to all those aged 9-55 years with at least three years of schooling and surveyed schools in the enumeration areas. Working with both GSS and the Ministry of Education, Youth and Sport (MOEYS), IEG resurveyed the same 85 communities and their schools in 2003, applying the same survey instruments as previously. In the interests of comparability, the same questions were kept, although additional ones were added pertaining to school management, as were two whole new questionnaires – a teacher questionnaire for five teachers at each school and a local language test in addition to the math and English tests. The study thus had a unique data set – not only could children's test scores be linked to both household and school characteristics, but this could be done

in a panel of communities over a fifteen year period. The test scores are directly comparable since exactly the same tests were used in 2003 as had been applied fifteen years earlier. The IOB evaluations relied on the use of secondary (administrative data), though they included some field work as well. The main sources are the annual school census data, national assessment tests, examination data, Demographic and Health Surveys (DHS), especially the EdData Surveys, the Population and Housing Census, SACMEQ II data¹⁹ and specific surveys for these studies (including a survey on teacher absenteeism).

Impact

The first major finding from the IEG study was the factual. Contrary to official statistics, enrolments in basic education have been rising steadily over the period. More strikingly still, learning outcomes have improved markedly. School quality has improved across the country, in poor and non-poor communities alike. Nevertheless, the differentials between both the poorest areas and other parts of the country, and between enrolments of the poor and non-poor, have been narrowed but are still present. Statistical analysis of the survey results showed the importance of building school infrastructure on enrolments. Building a school, and so reducing children's travel time, has a major impact on enrolments. Rehabilitation of classrooms can increase enrolments by as much as one third. Across the country as a whole, the changes in infrastructure quantity and quality have accounted for about one third of the increase between 1998 and 2003. Learning outcomes depend significantly on school quality, including textbook supply. Bank-financed textbook provision accounts for around one quarter of the observed improvement in test scores. But when satisfactory levels of inputs are reached — which is still far from the case for the many relatively deprived schools — future improvements could come from focusing on what happens in the classroom.

The IOB evaluations confirm the IEG findings for Uganda and Zambia. The examples show that a sector-wide approach can be an effective strategy for enhancing education within a relatively short period of time. The pooling of funds created the means for a broad, holistic and integral approach to the basic education sector. In both countries the strategy contributed to a large increase of enrolments. The education systems are still regressive, but differences tend to diminish. The econometric analyses confirm the results of the Ghana evaluation: investments in books, buildings and teachers had a significant (positive) effect on access and learning achievements. Nevertheless, while Uganda and Zambia succeeded in increasing enrolment, the quality of education remains low. In a way, the successful investment policy undermined its own success: it enabled the recruitment of more teachers and building of more classrooms to reduce pupil teacher

¹⁹ SACMEQ is the Southern and Eastern Africa Consortium for Monitoring Educational Quality. The consortium (of fifteen ministries of Education) tests the progress of pupils on reading and mathematics.

ratios and pupil classroom ratios, while at the same time attracting new entrants. Investments and the abolition of school fees had a positive impact on access, especially from vulnerable groups, and this resulted in higher pupil teacher ratios and higher pupil class ratios and had a negative impact on average learning achievements. The analyses also point to other problems. Teaching methods are old-fashioned and books are not always used effectively. High teacher and pupil absenteeism as well as high dropout rates (in Uganda) undermine the effectiveness of investments in the education sector. These weaknesses are related to severe underfunding, a lack of qualified and motivated teachers and head teachers and a lack of effective management capacity at the school and district levels. The analyses show that investments in schools, teachers, classrooms and books can be more effective when there is more attention for the governance and management of schools.

Impact evaluations of water and sanitation

Introduction

A lack of safe drinking water and sanitation facilities are important causes of diseases and (child) mortality. Especially diarrhoeal diseases are related with a lack of access to safe drinking water, in combination with a lack of sanitation facilities and unhygienic behaviour. A lack of access to water for productive purposes is also a key determinant of rural poverty and malnutrition (Rijsberman, 2004). Improved water and sanitation is one of the targets of the Millennium development Goals and central in many PRSPs.²⁰ Nevertheless, it proves to be more difficult to evaluate the impact of investments in water and sanitation. Many studies focus on health effects, but few impact evaluations look beyond health effects (IEG 2008, p. VI). A main problem is the availability of data. This note builds on the guidelines of the World Bank (PREM, 2006) for conducting an impact evaluation of water and sanitation, an overview of the IEG (2008) of impact evaluations of water supply and sanitation and experiences of the IOB (2007 and 2008c) of impact evaluations of the water and sanitation sector.²¹

Data and methodology

A typical impact evaluation of water and sanitation will use several data sources. When it comes to the output (such as water and sanitation facilities), the researcher may normally rely on administrative data. An assessment of outcomes is more complicated. In practice, many projects or programmes 'forget' to gather baseline data. As a result, it may be difficult to measure changes at the outcome level. Therefore, it appears to be important that an experienced evaluator is involved in the project from the start. In the case of an evaluation of the impact a large drinking water treatment plant in Sudan, the researchers could estimate the impact on beneficiaries using a simulation model, prepared for a baseline survey (IOB, 2009 (forthcoming)). Nevertheless, even in the absence of baseline data, it may be possible to conduct an impact evaluation. Especially at the sector level it may be possible to use information from household surveys (like the Living Standards and Monitoring Surveys (LSMS) and the Demographic and Health Surveys (DHS)). These household surveys contain information about water sources, the availability of sanitation facilities, knowledge of hygiene practices, acute and chronic health and demographic and socioeconomic characteristics (like income and education). Household surveys may be linked to administrative and census data. Moreover, it appears that information is not always available at the central level, but that there is much more information at the local level. For instance, for the IOB evaluation of water supply and

²⁰ Halve, by 2015, the proportion of people without sustainable access to safe drinking water and basic sanitation.

²¹ IOB is working on impact evaluations of water and sanitation in Tanzania, Yemen, Egypt, Mozambique and Benin. The Tanzania report has been published in 2007 and the Yemen report has been published in 2008.

sanitation programmes in Tanzania, the researchers did find important information about the prevalence of diseases at the local level. In many instances, additional surveys will be necessary.

The IOB evaluations on water and sanitation combine a descriptive approach (to establish the factual) with a quantitative impact analysis (to establish the counterfactual) and a qualitative evaluation of the (institutional) sustainability of the programmes. The impact evaluations combine existing administrative data with specific (household and community) surveys, carried out for the evaluation. Administrative data contain specific information about the investments in water and sanitation facilities, as well as information about education and health in the sampled communities. The surveys include information on water quality, diagnosis data recorded in health dispensaries and specific community characteristics (including information on water user groups). The analyses combine cross section and panel data with information from focus group discussions. The availability of panel data created the possibility of (first) difference regressions (or fixed effects regressions).

Impact

Impact evaluations of water and sanitation typically focus on (equitable and sustainable) access and use of water sources and sanitary facilities, hygiene awareness and practices, reduction of time spent on water collection and use of time savings (including school attendance), and health improvements. Observers may argue that it is not necessary to analyse health effects, as these effects are well known. The point is: we do not know. While it is true that the health effects of (enough) safe drinking water are known, this does not mean that we assume to know the impact of specific programmes. At point of use, water of improved sources may be of poor quality (IEG 2008, p. 9). The IOB evaluations, as well as other evaluations, do confirm this. The Tanzania evaluation showed how more than one million people got access to improved water facilities, realised by an investment of EUR 20 mln. Nevertheless, while the evaluation showed that improved facilities indeed contributed to a lower incidence of water related diseases, its effectiveness was sometimes undermined by contamination (at the source, during transport and transport at home), by the use of traditional drinking water facilities and by the lack of adequate hygiene training. The evaluation concluded that the quality of water, hygiene training and training of water user groups need more attention is needed. For Yemen, a country with much larger gender inequities, the researchers did find a large positive effect of the reduction of time spent on water fetching by women. Improved drinking water facilities had a significant (positive) effect on the school enrolment of girls. Zwane and Kremer (2007) did find comparable health effects for piped water and sanitation infrastructure and the private management of such services. However, they did find little evidence of the effectiveness of communal rural water supply infrastructure in fighting diarrheal diseases. Jalan and Ravallion (2003) and Galdo and Briceño (2004) did not find a significant impact of piped water on diarrhoea among the poorest groups, a

finding that may be explained by a lack of education (and training). These conclusions point to the same problem as mentioned by the researchers of the IOB evaluations: hardware facilities will only have a sustainable impact when the hygiene training and adequate management of the facilities have been guaranteed.

Impact evaluation of interventions to improve child health in Bangladesh²²

Introduction

Bangladesh might be one of the few countries that achieve the MDG of a two-thirds decline in under-five mortality by 2015. The improvements in child health (CH) over the past two decades are in excess of what can be explained by economic progress only.²³

Data and methodology

Amongst others, the evaluation measures the impact of publicly and externally supported programs on the improvement of child health since 1990.²⁴ The underlying theoretical model is based on a meta-review of empirical evidence on factors affecting child mortality. These include child-specific factors (e.g. gender and birth order), community-level factors (e.g. access to health care and sanitation), household-level socioeconomic factors (e.g. wealth and education), and unobserved factors (e.g. genetic family-specific characteristics).

The Demographic and Health Surveys 1993, 1996, and 1999 are used to identify the main determinants of CH in Bangladesh and establish the impact of related interventions.²⁵ Endogeneity of factors, such as antenatal care, is controlled for using instrumental variables. For example, the impact of antenatal care and delivery assistance on child mortality might be diluted because mothers who experienced complications during previous births are more likely to attend antenatal care *and* have children with higher risk of dying. Therefore, the equations rather include e.g. the presence of health and information facilities in the community and the distance from District health quarters, which are less likely to be correlated with child mortality. The relative importance of determinants of child mortality is assessed by their contribution to the decline in mortality during the 1990s and through examining the sources that explain the discrepancy between high and low mortality areas within the country.

The main determinants subsequently inform what interventions are covered in this impact evaluation. These are not limited to the health sector (immunization and training traditional birth attendants) but also cover the education (female secondary schooling) and infrastructure (electrification) sectors. Impact is measured by calculating the amount

22 OED, 2005, *Maintaining Momentum to 2015? An Impact Evaluation of Interventions to Improve Maternal and Child Health and Nutrition in Bangladesh*, Operations Evaluation Department, World Bank.

23 According to cross-country regression analysis, income growth accounts for at most 1/3 of the reduction in mortality between 1980 and 2000.

24 The report also covers fertility and aspects of maternal health, as well as nutrition.

25 CH is measured by the decline of under-five mortality, with different estimations for neonatal, postnatal and child mortality.

of under-five deaths averted per intervention on the basis of coefficients in the multivariate and the cross-country regression analysis. For example, the DHS analysis indicated that immunization reduces the probability of death by 50%, which can be used to calculate the deaths averted.

The impact of the interventions is subsequently combined with cost data in order to compare their relative cost-effectiveness. Attribution of the impact to each external donor (World Bank and DFID) is obtained by linking the number of deaths averted to financial contribution (omitting the impact of existing national systems and capacity). Moreover, this quantitative work is supplemented with qualitative analysis of the historical political context, changing role of women and private sector involvement.

Impact

A variety of factors have influenced the decline in under-five mortality in Bangladesh, but most notably economic well-being and social sector interventions. Immunization proves particularly cost-effective, saving up to 2 million children under 5 for less than US\$ 200 per death averted, but increased female secondary schooling and training of traditional birth attendants²⁶ are also important. Rural electrification reduces mortality through improved health services and access to health information from TV. Given that the World Bank and DFID support these programs, the evaluation concludes that external assistance has had a notable impact on child health outcomes in Bangladesh.

From the evaluation can also be learned that existing national surveys are useful for ex post impact evaluation if no evaluation framework existed at the start of the programme, though some adaptations might be required (e.g. more detailed community questionnaires).

²⁶ Interestingly, this conclusion comes after the program of training traditional birth attendance has been closed due to “faddism on the part of the international community, rather than a decision based on solid local evidence”, such as impact evaluation. OED, 2005, p24

The welfare impact of the World Bank's rural electrification programmes²⁷

Introduction

The World Bank's rural electrification (RE) program concentrated on outputs (coverage), assuming outcomes would follow automatically. As this connection can not be taken for granted, the Independent Evaluation Group (IEG) made a special effort to examine the social and economic impact of the RE lending portfolio in 10 different countries.

Data and methodology

The evaluation analyses the impact of RE on time use and the welfare implications of these changes for health, education, and microenterprise.²⁸ Using a theory-based approach, inputs and outputs are identified and outcomes are linked to those outputs and who receives them. Because eighty percent of total rural electricity consumption is used for lighting and television (outputs), this is the main focus of the evaluation.²⁹

With regards to health, the evaluation measures

- a. how access to information from TV and radio affects health knowledge;
- b. how knowledge changes health behaviour and;
- c. how this in turn impacts on health outcomes and fertility.

The evaluation addresses how electricity affects education by

- a. improving the quality of schools; and
- b. increasing study time.

The impact of RE on microenterprise is measured through:

- a. availability of complementary infrastructure (e.g. roads, bank);
- b. stock of equipment of enterprises;
- c. hours of operation.

The evaluation uses data from Demographic and Health Surveys (health and family planning), and household Income and Expenditure Surveys / Living Standard Measurement Surveys (income generation). Most of this is cross-sectional data, which

27 IEG, *The Welfare Impact of Rural Electrification: A Reassessment of the Costs and Benefits*, IEG Impact Evaluation Series, 2008

28 The report also addresses the contribution of the WB to increased coverage of RE, the distributional profile of those taking connections, the unit costs of connection for users and suppliers, as well as the willingness to pay for connection (which overlaps with the welfare impact).

29 As electricity is only rarely used for cooking (though a bit for rice in East Asia), the potential benefits from displacing firewood or kerosene stoves are limited (e.g. health benefits from improved indoor air quality or global benefits such as reduced carbon dioxide emissions). These impacts are therefore not quantified in the evaluation.

compares households at a certain moment in time, though for two countries panel data was also available to analyse changes over time.

As with most impact evaluations, this study risks selection bias, given that the beneficiaries are not a random sample of the population. The beneficiaries may differ significantly from the (counterfactual) group that has no access to RE. For example, in many countries communities are connected on a “least cost” basis, which favours communities in areas where the costs of electrification are relatively low (i.e. less remote areas). The characteristics of these communities may differ from the communities in more remote areas (for instance with respect to income and educational background). Nevertheless, when the selection criteria (such as location) are observable, a regression-based approach, may overcome this bias by including these criteria in the model. As a result, the analysis captures the effect of electrification compared with the counterfactual of no electrification. The impact of electrification on microenterprise by households was analysed using a difference-in-difference approach. This procedure ensures the elimination of an (initial) selection bias by capturing changes over time rather than initial differences in characteristics of firms. The study also used a housing quality index as an instrument for income, because income may be endogenous (if electrification has an impact on income and income affects electrification).

Impact

Electrification extends waking hours, which are mainly used to watch television. Data confirms that electricity affects *health and family planning* knowledge through this increased access to TV. The link between health behaviour, measured by use of modern contraceptives and child immunization, and health knowledge is also confirmed by the data. However, the link between electrification and health outcomes, directly or through changed health knowledge and behaviour, yields less good results and requires more study. A health impact evaluation in Bangladesh showed a significant and direct impact of electrification on mortality.³⁰

The *education benefit* of RE is sizeable; children in electrified households have higher education levels than those without electricity, even when taking into account factors such as parental education and income. This might be explained by better teaching, as there is some evidence (Ghana) that availability of electricity makes rural positions more attractive for teachers. The significant direct impact in some countries could also be due to increase in reading and studying hours with illumination, though more time-use data is required to confirm this finding. RE also seems to indirectly improve the propensity of a child to stay in school via the mother’s knowledge and education.

30 IEG Impact Evaluation Series, 2005

In general, the impact of RE on productive activities is limited, unless when combined with programs to stimulate productive uses of electricity. The small but significant impact of electrification on the revenue of *home enterprises* is probably mainly due to increases in the number of hours worked per day and the use of electrical equipment.

Though this impact evaluation leads to some interesting new findings with regards to RE, the report also acknowledges weaknesses in the data, calling for more data collection specifically designed to examine impact.

ACRONYMS

ADB	Asian development Bank
AFD	French Development Agency
CGE	Computable General Equilibrium
CH	Child Health
DAC	Development Assistance Committee (OECD)
DFID	Department for International Development
DHS	Demographic and Health Surveys
EC	European Commission
GBS	General Budget Support
IADB	Inter-American Development Bank
IEG	Independent Evaluation Group
IOB	Policy and Operations Evaluation Department, Ministry of Foreign Affairs, NL
IP	Issue Paper
LSMS	Living Standards and Monitoring Surveys
MAMS	Maquette for MDG Simulations
MDGs	Millennium Development Goals
MOEYS	Ministry of education, youth and sport
NONIE	Network of Networks on Impact Evaluations
ODA	Official Development Assistance
OECD	Organization for Economic Cooperation and Development
OLS	Ordinary Least Squares
PAF	Performance Assessment Framework
PETS	Public Expenditure Tracking Survey
PFM	Public Financial Management
PRSP	Poverty Reduction Strategy Paper
QSDS	Quantitative Service Delivery Survey
RE	Rural Electrification
RHs	Representative Households
SBS	Sector Budget Support
SIE	Statistical Impact Evaluation