Conférences & Séminaires

05

Evaluation and its Discontents: Do We Learn from Experience in Development?

Proceedings of the 9th AFD-EUDN Conference, 2012

François BOURGUIGNON
Michael A. CLEMENS
James MIRRLEES
Jean-David NAUDET
Leonce NDIKUMANA
Jodi NELSON
Catherine PARADEISE
Ruerd RUBEN
Miguel SZÉKELY

Evaluation and its Discontents: Do We Learn from Experience in Development?

Proceedings of the 9th AFD-EUDN Conference, 2012

AUTHORS
François BOURGUIGNON
Michael A. CLEMENS
James MIRRLEES
Jean-David NAUDET[1]
Leonce NDIKUMANA
Jodi NELSON
Catherine PARADEISE
Ruerd RUBEN
Miguel SZÉKELY

[1] Jean-David Naudet's paper was co-authored with Tanguy Bernard and Jocelyne Delarue, AFD economists.

Conférences et séminaires

AFD's Research Department organises a large number of seminars and conferences which provide a forum for exchanges of knowledge and experience between development aid actors: researchers, experts, political leaders, NGOs, companies… These events may address any of AFD's fields of operation. The Conférences et séminaires *collection aims to provide readers concerned by these issues with the main outcomes and lessons-learned from this research.*

All our publications are available at http://recherche.afd.fr

[Disclaimer]

Publications Director:
Dov ZERAH

Editorial Director:
Alain HENRY

Designed and produced by Ferrari/Corporate – Tel.: 01 42 96 05 50 – J. Rouy/Coquelicot
Layout: Vif-Argent - Tel.: 01 60 70 02 70
Printed in France by: Imprimerie de Montligeon

Conférences & Séminaires

Table of Contents

Acknowledgements

This volume presents the proceedings of the conference, "Evaluation and its discontents: do we learn from experience in development?", jointly organised by the *Agence Française de Développement* (AFD) and the European Development Research Network (EUDN) and held in Paris on 26 March 2012[2].

For nine years now, AFD and the EUDN have teamed up for the annual edition of this major conference on development. Hosted this year in the conference centre of the French Ministry of the Economy, Finance and Industry, it attracted over one thousand participants from over thirty countries across Africa, Asia and America.

With the years, this event has become one of Europe's landmark meetings for the development community. Each year, it takes up the same challenge of bringing together a large, broad-based audience, including public institutions, businesses, NGOs, researchers, committed volunteers, etc., along with top experts who each make ground-breaking contributions to the chosen theme. The conference is, first and foremost, an opportunity for exchange. This has been the spirit of the previous editions, which have addressed major themes such as global inequality, international migration, the role of culture, managing natural resources under the pressure of population growth, the fragmentation of our globalised world, and measuring development.

In 2012, we saw fit to challenge a relapsing issue: our capacity to evaluate development and draw lessons from our experience in development. Are we willing and able to evaluate and account for development results and, if so, how is it that we seem unable to translate this experience into practice? What are the factors hampering the learning process?

We would like to thank the members of the EUDN for their fruitful collaboration and the high quality of their contributions. Our thanks also go to all the participants whose presence helped to make a success of this ninth joint AFD-EUDN conference. Last but not least, we wish to extend our thanks to Emilie Aberlen, Louis Blazy, Philippe Cabin, Sandrine Laborie, Véronique Sauvat and Laurence Wunderle for their precious help both in both organising the conference and editing this volume.

Robert PECCOUD
Director of Research
AFD (2002-2012)

[2] The proceedings of the previous EUDN conferences are available at http://recherche.afd.fr

Introduction

François Bourguignon, *Paris School of Economics*

On behalf of the European Development Network, let me also welcome you to this new ninth AFD-EUDN conference on development. This promises to be a great conference on a theme that is increasingly relevant and important for development policymaking – the evaluation of policies. There are many reasons why policy evaluation has become so vital and yet so elusive in today's world, and some of these were mentioned in the introduction to the conference. I am sure that our discussions today will address this topic under various lights. To my mind, the current focus on the results of policy evaluation derives from the conjunction of three trends. First, there is an increasing social demand for greater accountability from policy makers, bureaucrats and NGO managers towards their constituencies, the taxpayers, contributors and the intended or actual beneficiaries of specific programmes or policies. This demand may reflect the robust progress being made with respect to the conduct of democracies and the growing demand for transparency. Second, there is a demand for more efficient public spending, combined with the increasing tendency to use market instruments and market values in all spheres of public intervention, In the same vein, the bias towards greater quantification in public – and thus in non-market – activities also explains, to my mind, this increasing demand for evaluation. Third, there is the widely shared view that, particularly in the development field, efficiency requires learning as much as possible from the experiences of others in the same country or further afield. Yet learning is practically impossible without reasonable evaluation of the actual development impact of specific projects, programmes and policies. In one way or another, the ideal world that the actors of public life have in mind is a world where it could be said that one euro spent on a given policy or programme yields the equivalent of x euros in social benefits, whereas another programme would yield y. Such quantification of the costs and benefits of programmes and policies would clearly allow us to determine which raft of programmes could best be implemented for a given availability of funds, to judge whether the choice of policies by policy makers or NGO managers was appropriate and, for benchmarking purposes, to examine whether projects, programmes and policies have been conducted effectively.

In the field of development, the above description of policy evaluation is only a slight exaggeration of the demand made on national and international development agencies, policy makers in both donor and recipient countries and NGOs to scrutinise the effectiveness of development aid. I remember very well a member of a European government coming to ask me, when I was in the World Bank, what the overall rate of return was on the development aid that his country was contributing to the World Bank. And I have heard, or seen, many requests of this kind on other occasions. Of course, what a wonderful world it would be where such simple quantitative evaluation were possible, but also how dreadful it would be if such simple quantitative evaluation were common practice. The world is much too complex for such cost-benefit analyses to be possible on a systematic basis. And, I would add, from some points of view, it is very fortunate that things are so.

The difficulties of evaluating specific programmes in the real world are indeed numerous and often almost insurmountable. It is difficult to figure out ex ante what the outcomes of a project will be. It is equally difficult task ex post, except within some specific kinds of experimental frameworks. In all cases, it is difficult to evaluate outcomes in terms of the final social objectives. More specifically, it is difficult to aggregate the various dimensions, especially the non-economic dimensions, of those objectives. Finally, it is difficult to assess the precise role of the context of a programme or policy in the final result, and therefore to draw some generally applicable conclusions from one specific evaluation.

The title of this conference echoes Freud's well-known book, Civilization and its Discontents. I know that some of you were somewhat shocked by this title, which is seemingly a strong critique of the evaluation work carried out these days. I think that the title of Freud's book in its English and French translation is not completely accurate, not completely satisfactory. In our development context, and because of the difficulties I just mentioned, I believe that the German word "das Unbehagen" would be more aptly translated by "uneasiness", the problem being that the English term "uneasiness" is very difficult to translate into French. So the best solution seemed to be to stick to the official French and English titles of the book. But I hope that with this clarification, you understand exactly what the objective of this conference really is: to make evaluation work and all the discussions about evaluation methodologies a little easier to handle. And, of course, various methodologies have been proposed to deal with all the difficulties I have just mentioned. Some of them will be discussed today, with special emphasis on the experimental approach, which has attracted a great deal of attention over the last ten years, and on indicator-based management, which may be viewed as some kind of integration of evaluation and management objectives. Nowadays, this is a virtually generalised practice in business enterprises as well as in government, international agencies, bilateral agencies and private foundations. Of course, the discussion will cover other methodological aspects of evaluation, especially when dealing with the way in which it can actually help us to learn about development experiences so as to better design and monitor development strategies, including of course in the field of development aid.

Well, we have a wonderful group of prominent figures with us today to help us see more clearly how evaluation work in the development field can be thought out and organised. I would like to thank all of them for having accepted to contribute to our reflection today with their own experiences and thinking. The record number of attendees at this conference shows how important the issue of evaluation is to them and I thank them for their interest. Finally, I would like to thank AFD for its continuing collaboration with the development economists in the European Development Research Network. It is not so often that this kind of collaboration between developers, top development managers and operators, as are found in AFD, and academics takes place on such a scale and with such intensity. And we are very grateful to Mr Zerah and the whole AFD team for making this possible. Within AFD, we should also express our utmost gratitude to the pillar of this event, the mastermind behind this conference, Robert Peccoud. He has been the instigator of these AFD-EUDN conferences since the beginning and he is responsible for their growing success, very much due to his and his team's professionalism in organising these conferences not only on the logistics side, but above all with respect to their intellectual content. Robert will be retiring at the end of this year and I believe he would like this conference to be part of his swan song. So let's make sure that this day will match up to his expectation. Thank you very much.

Conférences & Séminaires

Part 1: Evaluating Development Policies

1. Evaluating Development Policies

James Mirrlees,Chinese University of Hong Kong

Abstract

This paper focuses on the issues at stake and difficulties involved in the evaluation of macro-level policies. Although evaluation is a complex area where consensus is hard to reach, it nonetheless remains an indispensible tool that is largely recognised as necessary to enlighten decision-making. From the perspective of "macro"-level evaluation, decisions that need to be informed are those conducive to growth, and GDP is used as the main yardstick to evaluate growth. The paper argues that although it is easy to use, GDP nonetheless entails a host of both technical and conceptual problems that lead it to become a rather unsatisfactory way of evaluating what the economy produces. Other determinants of growth (such as human capital, environmental costs, the value of the expertise required to evaluate) would need to be better taken into account, on the investment side, to provide a more faithful measure of the economic value of policies. The paper concludes on the other uses that can be made of evaluation, and recalls how evaluation can also be an efficient tool for assessing and rewarding the performance of managers and decision-makers to encourage the achievement of expected development outcomes.

1.1. Evaluation and Policy

We evaluate economic outcomes for the sake of economic policy. The policy might be very grand, such as entering the World Trade Organization or introducing widespread privatisation of production. It might be of medium scale. Changing the rate of taxation on company profits has macroeconomic consequences. One should also evaluate the probable consequences before deciding whether to build a new road or power station. In all cases, evaluation is not just describing the outcomes: it requires a judgment of value. The judgment may not be expressed in one dimension, and may not be numerical; but it could be, and perhaps should be.

While people can and do disagree about observations of economic facts, usually because questions about economic circumstances may be difficult to interpret, it should always be possible, in principle, to resolve the disagreement. Observations of a week's income, or of somebody's wellbeing, are standard examples. Disagreements about evaluation are inevitable, and may be impossible to resolve. Different people have adopted different principles. Yet there is considerable agreement on moral values. Persuasion and compromise are possible. In one sense, evaluation is impossible. But it should be done. The alternative of choosing policy blind is intolerable.

First, I will discuss the large-scale issues, those that we had hoped to resolve by observing the impact of institutional or macro-policy reforms on aggregate economic performance. Policy makers and economists converged on the question of what we must do to grow fast, as the basis for choosing these large policies. While this approach drastically simplifies and distorts the evaluation issue, it does frame and focus the issues quite helpfully. And I shall want to explain why I do not think that policy questions should be confined to these large-scale decisions about the economic system.

Having sketched a picture of economic development, I will go on to examine how it might be evaluated in ways more subtle and useful than by the growth rate of GDP. I want to reflect again on the many important aspects of human wellbeing that are neglected when we measure and compare GDP, and see what can be done to let them influence our evaluations.

That will lead us on to the micro-economic issue of project appraisal. So many development projects and programmes are hard to assess with a monetary measuring rod that it is not surprising that the influence of cost-benefit analysis on economic decisions has been less, perhaps much less, than we might have expected. Is that inevitable, or can quantitative evaluation, and its concomitant project design, contribute more to policy decisions? I will also try to say something about the fundamental problem of dealing with the costs of evaluation and decision itself: to put it simply – how many experts should there be?

Consideration of project evaluation will lead me to my final topic: the relationship between evaluation and incentives, an issue that seems to deserve much more attention than it has yet received. Of course, there are no easy answers.

1.2. What makes growth?

The measurement of GDP has been standardised to a degree, and we have data, or at least serious estimates, for quite long periods for every country. Figures are also quite up to date. We therefore know that a number of countries have grown rapidly, in the sense of GDP growth, over relatively long periods. By that measure, several Asian countries have grown about tenfold in fifty years. It is not mysterious how they have done it. Investment relative to GDP was high, and diminishing returns to capital growth could be avoided by drawing labour from low-productivity parts of the economy, mainly from rural areas. This is the model put forward by Arthur Lewis in the fifties, before much of the fast growth had happened.

My claim, put so baldly, that investment – in the form of investment projects – causes growth, is vulnerable to the objection that investment itself is endogenous, that is to say, is itself the result of many causes, among them decisions to create one production facility rather than another, and decisions by people who finance the project as to how much to save or lend. And, it will be said that these are what policy should attempt to influence. That is quite right. It is what we learn if we identify investment as the immediate cause of growth. And we must recognise that it is both the quantity and quality of investment that influences growth. Therefore, evaluating investment projects is not a simple matter.

Investment, at a sufficient rate and of adequate quality, makes growth possible, not certain. Other countries with low productivity and adequate saving cannot necessarily follow the lead of the Asian economies, though more and more do. So many circumstances can interfere with the process. Conflict., for one, is a major problem, as violence can prevent growth and bring famine. Accumulated wealth confined to the few can disappear into Swiss bank accounts. Capital can be invested in essentially unproductive technologies. And there is always the possibility, now so clear in the developed part of the world economy, that demand will be insufficient to bring production possibilities to fruition, discouraging further investment. We should not expect, when comparing countries and measuring their experience, to discover some simple formula relating growth of the economy to investment and other obvious variables.

1. 3. Production possibilities

What then does econometrics add to that crude and simple account of growth? There is plenty of data on GDP, investment and other variables for practically all countries over quite long periods of time. Many economists have used statistical techniques to deduce production possibilities from these data. One of the lessons seemed to be that investment is not as important as casual observation of the world suggests. I am not

sure, though, that the methodology of these studies is correct. Brock and Durlauf[3] have argued that it is wrong to include all countries in one large statistical analysis. It would be better to study groups of similar countries separately. That is already clear when one thinks about the Lewis model for developing economies, with substantial labour migration from old rural production to modern urban production, suggesting a very different picture from that for developed countries. The East Asian economies in their fast-growing phases might be grouped together. Yet they were at the various stages of growth at different times, with different technologies available, and followed rather different production strategies. It would be difficult to make a solid case for grouping them together.

Yet, I suggest that there is a more fundamental problem. The data for a particular country in a particular year tell us how much could be produced with capital and labour then and there. Comparing that with production, capital and labour in another year tells us very little about how much more production would have been possible that year if there had been more capital. Fitting a particular form of production function to the data, using statistical techniques, amounts to assuming without evidence connections between production possibilities at different times and in different circumstances. It should not surprise us if the results of the statistical fitting are often hard to believe, and should not be believed.

Really, it is the fault of history. The world has, not provided enough natural experiments. We should not have expected that a wide range of alternative inputs would have been tried just to show us what production possibilities were available. The data simply do not reveal with any precision or reliability what would have happened if investment had been different. The problems are compounded by errors in measuring capital and output.

In fact, for the Asian economies, when these experience fast growth, the ratio of each year's increase in GDP to investment in the year before, has remained fairly constant, at least until the recent crisis. It is only suggestive, but the relationship supports the claim that investment causes growth. A country grows faster when it invests more. These results are consistent with the ideas suggested by the Lewis model: that it is particularly investment that makes growth possible, and that, in order to achieve fast growth, there should be considerable initial inefficiency in the economy. There are substantial differences among countries, which presumably show the influence of good investment decisions, and good economic environments, on GDP.

[3] Brock, W. and S. Durlauf, (2001a), "Growth Empirics and Reality," *World Bank Economic Review*, 15, 2, 229-272.

1. 4. Other contributions to growth

We should rehearse the other elements of the economic landscape that contribute to growth and interact with capital investment. Many of the policies that need to be evaluated are intended to influence one or other of these other factors.

Human capital is always the first to be mentioned. The ability of people to work, to do specialised tasks and to take good decisions is influenced by education, experience, particularly at work, and by healthcare. There is evidence of a considerable impact of even quite simple medical procedures on future earnings; and, of course, people with more years at school on average earn more too. The higher earnings usually mean that they are responsible for more of GDP. Projects to improve and expand education, projects to provide medical care and projects to provide work experience all have this capital aspect, as well as providing the immediate pleasures of learning, reduced pain and contributing to society. Most of these effects are rather difficult to evaluate. Fortunately, experiments are providing plenty of good evidence. Abhijit Banerjee and Esther Duflo[4] mention several of them in their fine book, *Poor Economics*.

Many economists and most people regard international trade as a major contributor to economic growth. International trade allows a country to exploit economies of scale and to specialise in producing what its capital stock and labour force are best suited to produce. In effect, most countries that want an aeroplane can produce it by growing cotton, making textiles and trading the appropriate amount for the plane. The value of what is produced is enhanced by international trade opportunities. So simple a point shows clearly why in evaluating output (or inputs) one should use world prices to evaluate them.

For most producers, other than subsistence farmers, getting their product to market is a major concern, and the cost of doing so may well be a considerable part of the price to final consumers. The availability of roads, docks, airports, electricity, drainage and the like can clearly have a great influence on production opportunities. The legal and security framework of the country obviously matters too, with all it implies for the cost of contract enforcement, corruption costs, compliance costs and various kinds of protection.

This is all to recognise that evaluating policies intended to produce growth is not simply a matter of measuring the cost of investment and estimating the additional output that will follow.

[4] A. Banerjee and E. Duflo (2011), *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, PublicAffairs.

1.5. Decisions

Economic policies are the result of decisions, and they are usually intended to influence decisions. The decisions will be based on evaluations of the consequences of alternative choices. I do not say that they will necessarily maximise some number representing the evaluations. The choice made usually depends on a less formal response to evaluations, and the evaluations themselves may well not be expressed numerically. But decisions have to be made, and the simplest way of modelling them is to think of maximising the net value of the consequences, even if we think it will be a very rough and ready kind of maximisation.

Decisions relevant to the economy take many forms. First and most important is choosing (and designing) projects. That is not all that an economic manager has to do. He has to find competent assistants and delegates. He has to consider directions of enquiry, research and development. In fact, many people have to make research decisions, large or small. Every decision depends on prior decisions about the gathering of information, by dumb search, by thinking or by designed experiment. That is what experts, advisers and consultants are for. We must think about how their contribution is to be valued, and chosen.

I suppose that the contribution of the development of contacts and connections would not be considered as important or fundamental to production decisions as the contribution of scientific and technological discoveries. But it is clearly highly valued, if we are to judge by the amounts businesses are prepared to pay for conferences and executive courses, not to speak of the fees at the leading business schools. One would like to think it is getting cheaper with the spread of social networking and news feeds, but it is clearly valuable. Again I am signalling a need to consider how to value expertise, which is a major part of development aid expenditures.

1.6. Evaluating the macro outcome

Before we descend to that level of detail, we need to acknowledge and respond to a big issue. GDP is a most unsatisfactory way of evaluating what the economy produces. If in some sense we want to estimate the human welfare generated by the economy in a year, then surely no one thinks that GDP is the answer. The only reason it is used so much, and implicit in every reference to economic growth, must be its easy availability. Another reason is that, as soon as you try to do a better job, you find all kinds of difficulties in your path: not just the difficulty of devising a measure or measures that everyone will accept, but the conceptual difficulty of just what is to be measured.

Is it the good being done by the economy, to people, in that year; or is it the good the output of the economy that year will do to

people, sooner or later? I suppose it should be the latter, if we want to judge the economy, or estimate the effect of large decisions on our valuation of economic output. That means we should estimate the contribution of current investment to future consumption and other goods. It also means we should estimate the impact of health expenditures, and other relevant consumption, this year on future life and death. Current life expectancy (as used in UNIDO's Human Development Index) does not do that directly. It is a kind of proxy, but we need to think about how to do better if we are to measure such effects properly.

We could restrict our welfare measure to current consumer expenditure, private and public. That is certainly a contribution to human wellbeing. Investment may or may not contribute to consumption in the future. If markets are working properly, and the level of investment is not going to be too high in the future, the value of investment is the present value of the consumption that will be provided by it in future. But in fast-growing economies it is quite possible that part of investment expenditure only provides for future investment, in turn providing for further future investment, and so on until the world ends. That part of GDP is doing nothing for human welfare. But it is a matter of wild conjecture to determine how much. On balance, it seems best to accept GDP rather than consumption.

Why do we not offset GDP with the cost of the labour we put into it? Economists assume that people are paid to work because work is unpleasant, at least the amount of work that people are induced to do. If so, it seems we should deduct from GDP the total cost of labour, the wage bill, and calculate change in real income by valuing changes in production and changes in total hours of labour at the appropriate prices and wages. That would change comparisons of the USA to European countries, since hours are much higher in the USA. The big problem with this is involuntary unemployment, since being unemployed is generally agreed to be worse for most people than doing fifty hours a week; and that is confirmed by happiness studies. We would need to estimate hours of unemployment and multiply by a figure for the welfare cost, something we do not know how to estimate. Even if that could be done, it would suit the way we think about income better if we had a separate figure for labour costs rather than subtracting from GDP. Nevertheless, there is a strong case for subtracting an allowance for unemployment. The economy is supposed to allow people to contribute their labour to the common good. To the extent it does not do so, it should get less credit.

The contribution of the economy to health and education are, as everyone is aware, not well measured by expenditure on health care and schools. Many countries will try to estimate these contributions by results. This seems better than having separate measures of health, life-risks and literacy. The UNIDO Human Development Index in effect double counts, by using GDP figures that include education and health expenditures, and at the same time using years of schooling and life expectancy as proxies for results. Giving some weight to any information we have about effectiveness is of course admirable.

The economic value of things governments provide is usually not well measured by their cost to government. That is understandable, and not easy to correct. The problem of

evaluating public goods and environmental degradation would take us too far away from the main issues I want to address. But the issue is quite big. Measures of pollution, deforestation, water availability and changes in fish stocks subtracted from GDP would increase attention to increasingly serious problems. The ideal of a green national income figure deserves support.

1. 7. Who gets what?

The main flaw in GDP is that it takes no account of inequality. If we want to assess the value of economic activity this year, should we not give more weight to increases in low incomes than increases in high ones? Probably everyone would agree that the relief of poverty needs special weight. If that is agreed, does it not follow that income distribution matters? What can one reasonably do about it? One way is to include a measure of inequality in an index that also has the GDP. That is what UNIDO does in its Inequality-adjusted Human Development Index. They must have a great deal of difficulty in producing up-to-date figures, since information about income inequality, and particularly poverty, is normally available for low-income countries only many years later, if at all.

If it is important to be able to assess a country's economic performance taking account of who gets the benefits, then it is worth some considerable expenditure on obtaining up-to-date consumption and income sample surveys. But it will not be easy to get accurate data from properly representative population samples. It is therefore important also to try to estimate the errors of measurement and selection in household surveys. Without that, it would be difficult to try to correct for selection biases that seem to be growing over time.

While I think this UNIDO index is a good way of tackling the problem, I wonder if it might be worth considering a simple weighted average of people's incomes. The weight would be inversely related to the individual's income, for example one over the square of income. Individual income should also be adjusted for a person's age in some way. Then the change in GDP per head from the previous year would be calculated as the changes in income weighted appropriately. Changes in income for higher-income people would practically not count, and to that extent it may sound extreme; but it seems to me that it expresses values that are widely held, even by many of the rich. A crude and simple version of this idea would be to use a truncated national income that comprises the incomes of only the lower half or lower quarter of the population.

1. 8. Plural values?

It has been suggested by Amartya Sen that we should not confine our attention to a single number representing an economy's achievements, but should list several measures (not necessarily all numerical). They might measure private consumption, education, health, and inequality, and not be combined into a single index, since that index would be an arbitrary combination with no particular rationale.

The strength of the proposal to evaluate along several dimensions is that it avoids unreasoned balancing of different goods and bads. The difficulty is that it is not clear quite how these multiple values would be used in making judgments. It is not impossible, though. One might approve policy changes only if they were expected to bring about improvements in all these dimensions. For example, we could consider a principle of accepting projects only when they are profitable and do no environmental damage. An alternative is to accept proposals that improve on at least one dimension. But that seems less attractive.

I think most of us would be uneasy about the idea of using plural values primarily because we have no idea how that would enable us to select investment projects. Similarly, one cannot select people for admission to university without somehow combining the different elements of information about them. Merely providing numbers measuring different dimensions runs the risk that those who have to make assessments will, explicitly or implicitly, weight them equally. That could be worse than concentrating on one dimension alone. The important thing is to combine the measures in a way that is deliberate, systematic and has been thought about. If we have a formal way of using the measures together, it can be applied consistently, it can be criticised, and it can yield results. It is always worthwhile to try to reason about the trade-offs among different goods and bads. Health economists have been quite successful in doing just that, when comparing different medical procedures.

1. 9. Investment appraisal

When we come to choose projects, do we have to take all these considerations into account? There are standard projects of a kind easily described and understood by economists and accountants. These are industrial projects, including roads, ports, construction projects, and many others. They are just the kinds of projects that nowadays one might think would happen anyway, if they are worthwhile, and if loans are available on commercial terms. They might seem not to require evaluation. What is needed is to ensure that the right kind of credit is available, i.e. that capital markets work.

In practice, that is not quite right. The lender will have to evaluate the project in order to check the creditworthiness of the borrower.

Of course, there are still plenty of standard investment projects financed by the World Bank or other development agencies. The cost-benefit analysis that is done seems often to be rather unsophisticated. Rates of return are estimated in advance and measured in *ex-post* reports. Shadow prices are not used. Environmental effects are not usually priced into the project calculation, as far as I can see, although they can play an important part in the design of the project.

If prices in the economy are not seriously distorted, for example by tariffs, there is nothing serious to object to in that. The useful principle applies that producer prices are the appropriate prices to value the inputs and outputs of the project, so long as the country's tax system can be taken to be optimal. That is a strong assumption, expressing respect for the values and competence of the country's government.

Many projects are non-standard, and raise difficult problems of measurement and evaluation, particularly of their outputs. That is true of health and education projects. Many projects provide support to government, primarily by supplying experts in, say, taxation, statistics, or water supply. I doubt that anyone has applied cost-benefit analysis to them. Water, sanitation and environmental projects must be as hard to evaluate as health and education projects.

Why is thorough cost-benefit analysis, with a proper reference interest rate, and some use of shadow prices, not often found now (though the Asian Development Bank does make considerable use of it)? The obvious reason is that there are rather few of what I called standard projects. Development assistance has done more and more in these other fields where serious numerical evaluation is quite hard, unless experimental results are available.

Experiments can provide evidence of, for example, income gains in future life from health procedures or education of a particular kind. It is not feasible to carry out experiments in advance to evaluate every alternative project. Indeed it would seem that experiments are likely to be most valuable for post-mortem evaluations. The real point is to discover general connections that will let project designers estimate the effect of their own projects. These estimates will be quite inaccurate. It is important not to be discouraged by the uncertainties. It is surely important to hurry on with many similar projects whenever experiment shows good results. That means accepting evaluations with considerable uncertainty, and getting on with the projects rather than waiting for further evidence. Slow development is harmful. The poor cannot wait.

1. 10. Evaluating government support

A considerable amount of development assistance is spent on experts. Their primary purpose is, I take it, to set up systems for gathering information, to assist in taking decisions and to monitor projects. This shows that taking good decisions is quite expensive. Perhaps the issue of employing experts should be left to governments,

provided with some adequate amount of budgetary support intended for the purpose. Governments might be thought to have the right incentives to call them in when needed.

However there is considerable moral hazard, since we cannot tell whether the government will think hard before employing an expert. The theory of moral hazard says that incentives are not enough to get the right outcome. We need to develop a theory of the optimal expenditure on decision-taking. In a sense that is not strictly possible, since the answer would be a decision itself. But we might get close. It is a general problem in management. Following a suggestion of a successful businessman, I propose that experts should be appointed one at a time, initially with a wide range of tasks to be accomplished, further appointments being made only when there is clear evidence of overload. This does not solve the problem of whether one is needed in the first place. That decision must be based on past experience. It cannot be true that it is always worth spending more on getting more, and more accurate, information, any more than it can be true that it is always good to spend more on science. In the case of economic and administrative information, it is only possibly worthwhile if the information will be used in decision-taking.

1. 11. Incentives

Evaluations are not only for decision-taking. They can also be used as the basis of an incentive scheme. Rewards, such as bonuses, or penalties such as sacking, can be based on a post-mortem or even intermediate evaluation of the project.

In many projects, the responsible parties – the managers and senior staff – are not the beneficiaries. Teachers in a school system do not gain automatically if their former pupils get well-paid jobs, though they should get some pleasure. It may be good to introduce some incentives. They will need to be designed into the project. One advantage of systematic *ex-post* evaluation is that it makes suitable bonuses possible. Of course the bonuses are part of the cost of the project, and should appear in the full evaluation.

Banerjee and Duflo[5] emphasise that when tasks are defined in a project, they should be feasible for most of the participants. Similarly, when bonus systems are part of the project, they should be designed in such a way that some bonus is quite likely to be paid. Economic models tend to emphasise the direct reward compensating the agent for her effort. But such payments can also express appreciation, which in turn encourages good work by improving morale. There must be a payment or else appreciation will not be expressed.

[5] Ibid.

It is not impossible to apply the same idea to country aid. It may seem absurd to suggest that members of the government could receive a bonus to reward large reductions in poverty, say. But something like that does happen. The Nobel peace prize does it. It is worth remembering that bonuses do not have to be very expensive. Could there be a United Nations House of Lords, with honours distributed on the basis of exceptionally good country evaluations? Membership should not be based, I would hope, simply on the unweighted growth rate of GDP.

Conférences & Séminaires

Part 2: Impact Evaluations: a Tool for Accountability?

2. Impact Evaluations: a tool for accountability? Lessons from experience in AFD

Jean-David Naudet, Jocelyne Delarue and Tanguy Bernard, AFD

"I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail"
Abraham Maslow, *The Psychology of Science, 1966*

Abstract

This paper relates the Agence Française de Développement's experience in the area of impact evaluations. Our purpose is to assess to what extent such studies, when designed before actual programme implementation, can provide the type of summative evidence that donors still seek when promoting them. Specifically, we rely on three large-scale randomised control trials and scrutinise their capacity to answer questions about the programme's underlying impact "on whom", "on what" and "of what". We conclude that experimental studies should be promoted to clarify the "tunnel-type" issues characterised by a limited number of well-specified homogeneous inputs, a tried-and-tested process, a short and external-event-proof causal chain, a rapid and stable take-up rate, a high and stable level of participation, and a set of outcomes measurable in the short run. While a number of such issues exist and are well worth studying experimentally to inform future development policies, few development interventions satisfy the required conditions, and summative use thus remains limited.

2.1. Introduction

Since the mid-1990s, donor agencies have been increasingly concerned with demonstrating their capacities to improve the lives of the beneficiaries of their interventions. This concern is epitomised in the Paris Declaration on Aid Effectiveness, notably by the donors' commitment to shift the focus onto development results and their measurement (OECD, 2005). And today, despite considerable controversy over the attribution of results, most donor agency websites post assessments of the number of children sent to school, farmers trained, malaria-related deaths avoided and, sometimes, households lifted out of poverty.

Yet, by the early 2000s, it was increasingly recognised that the academic literature on the growth-aid nexus had for the most part failed to establish the type of causal relationships needed to assess development policies (e.g. Easterly, 2003). And at the micro-level, evaluations of specific interventions also lacked the "robustness" needed to deal with the different biases needing to accounted for when assessing an intervention's impacts on beneficiaries (e.g. Duflo and Kremer, 2003; Banerjee, 2007. Thus, donor debates over policies or projects were deemed by some as being akin to "ignorant armies clashing by night" (Pritchett, 2002): heated debates occurred without any firm evidence arguing for or against the likely final impact of a given intervention, and donors responded less to evidence than to political fashion (Deaton, 2007).

Led by a number of academics, impact evaluations (IEs), similar to those used in the medical field and based on comparisons of adequately defined treatment and control groups, have since been proposed as a means of reliably estimating causal relationships between interventions and their outcomes. With such robust and cumulative evidence, they promise to be useful for identifying the missing links between academic research and development practices, for promoting trial-and-error processes for policy-building where experimental projects are implemented prior to their scaling-up, and finally for delivering knowledge on "what works" in development policies (CGD, 2006). And overall, the donor community has largely followed this call by collectively contributing to dedicated international funds and using IEs to "strengthen [their] internal monitoring and evaluation systems", as recommended by a report from the Center for Global Development (CGD, ibid.).

Yet, it is fair to say that in a context where there is strong demand for results, one of the prime motivations of development agencies has been to use impact evaluations as a way of demonstrating their effectiveness and only secondarily to use them as a learning tool. For instance, a recent Overseas Development Institute (ODI) survey on impact evaluation production and use reports: "The available evidence does seem to point towards experimental IEs being commissioned in order to fulfil account-ability purposes" (Jones *et al.*, 2009: 8). And

in fact, most reference books continue to highlight the power of IEs to reinforce accountability for development institutions (Khandker *et al.*, 2010, Gertler *et al.*, 2010). Accountability, in turn, suggests that IE should provide a summative measure of impacts as for instance defined by the donor community itself: "Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended" (OECD/DAC 2002: 19).

While recognising the significant improvements that experimental methods have brought to empirical studies in the field of economic development, this paper suggests that donors' use of IE for direct accountability purposes may be limited. In particular, we posit that treatment-control methods are best suited to addressing so-called "tunnel" issues characterised by a clearly defined and stable "treatment", short and external-event-proof causal chains, and an effective impact of the intervention on a large proportion of the targeted population intervention. This ensures that IE effectively delivers responses to three questions: *the impact of what*, *the impact on what*, and *the impact on whom*? Yet, most development interventions do not satisfy such prerequisites, which thus sets limits on the use of IE for donors' accountability needs. In essence, while impact evaluations are well-suited to helping *understand* development processes and testing the mechanics linking an intervention to a given outcome, they are less often appropriate to the donors' needs for *measures* of their impacts in the field.[6]

The argument is illustrated by the *Agence Française de Développement*'s (AFD) own experience with impact evaluations. Following the Center for Global Development's call for more IE to be undertaken, AFD has piloted several such IEs of its operations in recent years, including two large-scale randomised control trials (RCTs) in the area of microcredit in Morocco and health insurance in Cambodia. While AFD's main purpose was to learn more about this new evaluation methodology, it also hoped to provide robust evidence of the impact of its interventions. Using the world-renowned PROGRESA impact evaluation as a benchmark, we assess here the extent to which AFD-supported IEs were able to provide summative assessments of the impact of these interventions on their targeted individuals.

The rest of the paper is organised as follows. Section 2.2 briefly describes the three studies underpinning our argument: PROGRESA, Al Amana and SKY. Section 2.3 discusses the extent to which the studies were capable of answering the questions: *the impact of what*, *on what*, and *on whom?* Section 2.4 identifies the commonalities and differences of these three examples with

[6] Our focus here is on the use of impact evaluations from a donor's perspective. We do not address the subject of impact evaluation, including randomised control trials (RCTs), from a methodological standpoint. This is already dealt with by a large body of literature (see, for instance, the Symposium on New Development Economics in *Economic and Political Weekly* edited by Ravi Kanbur (2005), a Boston Review book edited by Abhijit Banerjee (2007), the 2008 "*What Works in* Development" conference held at the Brookings Institution in Washington D.C., or the World Bank blog on impact evaluations (http://blogs.worldbank.org/impactevaluations/) for glimpses of ongoing debates).

respect to the IE design and characterises the types of "tunnel" issues that are the most appropriate for experimental studies. We conclude with a call for donors' continuous support for experimental studies, but for reasons related more to learning about policy mechanisms and policy-building through experimentation than to fulfilling accountability needs.

2.2. Three impact evaluation cases for study

Over the past six years, the *Agence Francaise de Developpement* has engaged in financing and piloting several IEs. The Agency's main objectives were not only to identify the results of the projects under investigation for accountability purposes, but also to assess the potential of these new evaluation tools to produce relevant knowledge for the institution's different needs. More specifically, AFD's work was geared towards a greater understanding and appropriation of results so as to promote their utilisation and dissemination.

This paper focuses on two large-scale RCTs performed at AFD. In fact, it is mostly the RCT approaches that have been promoted within the impact evaluation movement, and these now account for the vast majority of studies undertaken. The focus on these two IEs facilitates comparisons with the IE benchmark, PROGRESA, which is briefly described below.

2.2.1. PROGRESA conditional cash transfers

PROGRESA (now called "*Opportunidades*") is a Conditional Cash Transfer (CCT) programme that has been operating in Mexico since August 1997. A key feature of the programme is to provide families with cash transfers as an incentive for them to adopt behaviour towards their children that enhances human capital. Parents are thus required to send their children to school and take them to health centres. The condition attached to the transfer is thought to be a crucial component, as it is expected that simple income effects of transfers do not necessarily translate into more schooling for children. Importantly, these transfers were given to women on the grounds that women tend to use resources to improve nutrition, and it was also hoped that this would help to emancipate them from their husbands' authority. Finally, a number of supply-side actions were also undertaken to ensure that enough schools and clinics were available to meet the increased demand generated by the programme.

As such, the rationale underlying the PROGRESA programme was very simple and clear: address poverty in the short run while tackling poverty in the long run *via* investment in human capital. Most results were expected in the short run and were easily quantifiable in terms of school attendance or preventive health behaviour. Other expected outcomes, such as the empowerment of women, were less direct. They were expected in the longer run and measured through proxy indicators.

The PROGRESA IE started in 1998 and relied on a large-scale randomised experiment: a subset of the programme's targeted communities was phased in based on a random order. The IE was one of the first large RCTs to be carried out on a development intervention, relying on a

sample of 24,000 households, and was thus highly publicised. Nearly all of the households that were offered the CCT took it up. Within eighteen months, results showed that the programme had a number of positive impacts on health outcomes and school attendance (for a review, see Skoufias, 2000).

PROGRESA is a key milestone for the IE movement owing to both the rigour of its methodology and the effects of the evaluation itself. In light of the results obtained, the programme was continued and expanded to urban areas, even though in the meantime there had been a change of government. Further, it has provided a model underpinning an extension of CCT programmes across the world.[7] Most of these have also included an IE – thus providing the necessary basis for the "robust" kind of meta-analyses targeted by the IE movement (Fizbein and Schady, 2009).

2.2.2. Al Amana rural microfinance

Al Amana is Morocco's largest microfinance institution, with more than 450,000 active clients in 2008. Since its inception in 1997, Al Amana, and other Moroccan microfinance institutions, have concentrated their activities in urban areas. Since 2006, however, Al Amana has been developing its client base in remote rural areas. Taking advantage of the fact that microcredit was still absent in these regions, an IE study was planned in order to compare households with and households without access to microfinance in the following years. Linking up with international debates on the capacity of microfinance to pull households out of poverty, the study sought to measure the impact of microcredit on levels of household income and consumption-expenditure (see Crépon *et al.*, 2006).[8] The study was thus meant to fill a large knowledge gap, being the first "robust" impact evaluation of microcredit in rural areas.[9]

The study design was based on pairs of geographically close and "similar" villages, from which one was randomly selected for immediate access to microcredit, while the other village served as a control village for two years. A total of 88 pairs of villages were selected all over Morocco to ensure the representativity of results on a national scale. About 6,000 households were to be surveyed at three points in time (before Al Amana began its activities, a year later and two years later).

A number of difficulties and changes of direction arose during the course of the project and research process. The most important issue concerned the take-up rate. Initial expectations were mostly grounded on Al Amana's own experience of its activities in urban areas, and rural participation was expected to reach similar

[7] In 2009, 28 such programmes were implemented in developing countries, compared to only three programmes ten years earlier.

[8] Other outcomes such as women's empowerment and children's education were also planned as secondary questions.

[9] The study design largely followed that of another impact study of microfinance in Hyderabad, India. Other such impact studies at the time were in preparation, including the one in Hyderabad (see Banerjee *et al*. 2010) and another in the Philippines (see Karlan and Zinman, 2010).

levels. On this basis, a conservative estimate of a 60% participation rate within the general population would make it possible to detect only a 20% change in the final consumption level (Crépon *et al*., 2006). In order to enable detection of smaller effects, the sampling scheme selected the 25 households per village that had the highest probability of becoming borrowers under the newly set up microcredit scheme, based on a propensity score built using data collected in the feasibility study (Crepon *et al*., 2007). It soon became clear, however, that the take-up rate was much lower than expected.[10] Several measures were thus taken to encourage higher participation: the midline survey was cancelled to give enough time for take-up rates to rise and some of the product's features were changed (removal of quotas for women, modification of repayment schedule, better information available to the population).

The results of the study show that, given sample size and take-up rates, no impact was observed on poverty, consumption, activity diversification or shock absorption, although some significant effects on production and wages were identified, particularly for those households above the median poverty level (see Crépon *et al*., 2010*a*).

2.2.3. SKY health insurance

SKY ("*Sokhapheap Krousat Yeung*" signifying "Health for Our Families" in Khmer) is an innovative micro health insurance programme operating in Cambodia. SKY was created by the French NGO *Groupe de Recherche et d'Echanges Technologiques* (GRET) and aims to improve the health of Cambodians by providing affordable health insurance and quality healthcare without the risk of impoverishment. For a fixed monthly premium, SKY offers households free and unlimited primary and emergency care at contracted public health facilities, as well as a number of other services. By 2008, SKY was operating in four provinces (Takeo, Kandal, Kampong Thom and Kompot) and in the capital, Phnom Penh (see Levine, 2010, for a detailed description).

The SKY IE uses a randomised control trial to examine the causal effect of the proposed insurance on households' economic and health outcomes and on their decisions regarding healthcare utilisation. Its also aims to understand who does and who does not choose to purchase insurance – in particular, to identify the issues of adverse selection that are omnipresent in the insurance literature. To this end, the study relies on the random allocation of discount coupons for the purchase of a six-month insurance coverage. This arrangement enables the comparison of those households that did contract insurance using the coupon with similar households that did not contract insurance but would have done so had they been given a coupon.

Here also, a number of difficulties arose. Drop-out rates were significant once the discount period had expired, which led to an extension of the coupon scheme. In addition, the study aimed to assess how insurance would affect a household's debt burden

[10] This averaged 17.6% after 24 months of programme availability to eligible individuals according to Al Amana's administrative data, 10.6% according to the survey data, and 13.6% among the 25 households per village predicted to be the most likely to borrow (Crépon *et al*., 2010*b*).

following a severe health shock, which is an event too rare to detect statistically. Finally, the information gathered on health services consumption was too parsimonious to derive any meaningful measure of changes in behaviours, and little improvement in health outcomes was identified. The study does however identify some economic improvement in the life of those households where one of the members fell ill. Finally, the evaluation uncovers some signs that adverse selection had influenced participation in the insurance scheme.

2.3. Towards accountability? Summative impact evaluations

We use the three studies described above to illustrate the potential of IE to provide summative assessments of a programme's impact. In a nutshell, summative evaluation implies that the study is capable of assessing a programme's overall impact under (close to) normal conditions. More specifically, we use these three studies to provide some elements of an answer to three different questions about what kind of impact is in fact being measured: The impact on whom? The impact of what? The impact on what?

2.3.1. On whom is impact measured?

The impact evaluations rely on sampled units, which may or may not be representative of the underlying population of beneficiaries. Here, a key factor is the participation rate. Quite logically, only those units that participate in the programme's treatment group and those individuals from the control group who would have participated had they been offered the programme enable the effect of the intervention on its beneficiaries to be estimated. However, in cases where participation is not mandatory, and thus the participation rate is not 100%, it is not possible *ex ante* to distinguish with any certainty those units that will participate from those that will not. This means that the best way to draw a sample nonetheless representative of all participants is to randomly select units from the pool of all targeted beneficiaries.

If the participation rate is high, most of the randomly drawn samples will be useful for the estimation. This is the case of PROGRESA, which has a participation rate of more than 98%. If the participation rate is limited, however, only a fraction of the sample will be used to estimate the impact, which in turn could limit its statistical power to detect an impact of reasonable magnitude. It is sometimes argued that one could still assess the impact of the intervention on the targeted population – rather than only on the actual participants – by taking into account eventual spillover effects from participants to non-participants within the targeted population. And in fact, this estimator – the intention-to-treat estimator – may be more relevant from a donor's perspective as it comes up with an answer to questions such as to what extent has well-being in this region been changed as a whole thanks to the intervention (Ravallion, 2008, pp 36-37). Yet, unless spillover effects are strong and widely disseminated within the targeted population, it is unlikely that the IE will be able to capture them if participation rates are low.

To take this participation issue into account, a very purposive sample selection scheme was implemented in the Al Amana study (cf. Section 2.2.2), whereby only those individuals who were most likely to borrow

would be included in the sample. In cases where the predicted probabilities included the vast majority of actual borrowers within the sample, the resulting impact estimates would, in effect, be representative of the total population of borrowers. Yet, the models used for purposive sampling most often have positive but limited predictive capacities and are only capable of marginally increasing the sample's participation rate. In the case of Al Amana, the borrowing rate in the targeted population was 11%, compared with 13% in the sampled population. Overall, the study's population is clearly not representative of either the population of targeted households or the actual participants. Furthermore, the complexity of the model used for prediction limits the study's capacity to replace sampled individuals within the general targeted population. As a consequence, the evaluation informs us on the (non-) existence of an impact and its magnitude, but only for a very specific and purposely built sample of the population targeted. In effect, it fails to provide a measurement of the programme's impact on a sample representative of the targeted beneficiaries, which is the key information needed for accountability purposes.

The problem is slightly different in the case of SKY, where village-level randomisation was not feasible for mostly operational reasons and the random exclusion of households within the community was deemed unfeasible for ethical reasons. This was overcome by randomly allocating, within villages, discounts on the purchase of several months of insurance, which in turn caused an exogenous change in the probability that some households would participate while others would wait. A Local Average Treatment Effect estimator can then be computed and would enable estimation of the impact of insurance within the population of households for whom the discount did make a difference in the decision to participate or not (the "compliers").

Here again, however, it is worth questioning to what extent the population on which the treatment is estimated is representative of the population for which the programme is normally implemented. And in fact, by design, impact is assessed in this case on a population that would not have chosen to participate had the insurance been priced at normal rates. In fact, the "always-takers", or in other words those households that would participate under normal conditions, do not contribute here to the impact measure – this is akin to what is sometimes referred to as "randomisation bias" (e.g. Ravallion, 2008). Overall, it is quite likely that the impact on those who are willing to purchase the insurance at a higher price is different from the impact on those who only join when it is discounted.[11] Consequently, the magnitude of aggregate causal effects on all beneficiaries of the insurance remains unknown at the end of the evaluation. Here

[11] If one wanted to measure the impact on those who participate under normal conditions, the best thing to do would be to provide negative incentives to some (such as higher prices than normal), and normal conditions for others. In this case, the pool of compliers is likely to be representative of those who participate under normal conditions. Obviously, however, making people pay more than they would under normal conditions for services that we know are potentially beneficial, such as health insurance, raises ethical issues.

again, a key piece of information is thus missing if the study is to be used for accountability needs.

Overall, while in the case of PROGRESA the study sample was most likely representative of both the targeted and the actual beneficiaries, this is less the case for Al Amana and SKY. In these studies, the responses to weak take-up rates and the need to randomise at an individual level produced purposive or partial samples no longer representative of the populations that are relevant for donors' accountability needs (be it the targeted population or the population of actual participants).

2.3.2. The impact of what is measured?

We now turn to the projects evaluated in order to investigate how their evaluations can be used for accountability purposes. In fact, we shall see that the actual impact measured by the IE may not be as easily interpretable as the impact of the projects as we need for accountability purposes.

We first ask whether the project implemented can be interpreted as if it were being implemented under "normal" conditions. In the case of PROGRESA, the IE only examined a subset of the communities targeted by the Programme, while at the same time PROGRESA was being rolled out throughout the country using the same design.[12] Overall, one can feel relatively confident that the IE produced results are very much representative of the programme as implemented under normal conditions.

The situation was different in the case of Al Amana, where the programme was in the main implemented without due thought being given as to how to adapt it to the new rural clientele. In fact, Al Amana had thus far never operated in the targeted remote rural area, and a number of issues involving the programme itself had not been addressed. Al Amana had decided to first implement the scheme on the same lines as those used in urban areas and then modify it on a learning-by-experience basis.

For instance, the repayment schedules initially mirrored those used by Al Amana for microcredit in urban areas, with no consideration for the agricultural calendar. Later during the evaluation, however, some loans were made available that did not require repayment during the first few months. Also, quotas for women were initially implemented so as to encourage their participation but were abandoned as the study progressed due to the low take-up rate observed. Similarly, while it was initially planned for loans to be provided on a group liability basis, this later changed to allow for loans to individuals. Finally, despite the original decision to provide households with a normal level of product information, the weak take-up rate also prompted Al Amana to provide extra incentives to its field agents for the duration of the study.

In sum, the design of the Al Amana project had not been stabilised during the study period, and was very much considered by the Al Amana Morocco Credit Association as an ongoing learning experience. Since the IE

[12] See Barham, 2005, for an in-depth description of PROGRESA's roll-out.

had to be implemented at the same time as the opening of new branches, this inevitably involved evaluating a product subject to a high degree of change. As such, although the study aimed to determine the overall impact of access to Al Amana's microcredit scheme, this impact must be re-defined as the impact of access to a non-stabilised and hence not fully fledged intervention.

In the case of SKY, however, the project had already been piloted for several years in other communities, which meant that the intervention was henceforth mature and stable. The IE could thus effectively investigate the impact of the programme as normally implemented. Yet, a caveat must be added insofar as what was evaluated was no longer the impact of insurance, but rather the impact of an *almost free* insurance product designed for the purpose of the study. This is likely to be of consequence, since offering such a product may reasonably be expected to influence the beneficiaries' reactions. In a sense, the question addressed by the evaluation is the effect of "being offered what amounts to (almost) free insurance" and not the effect of "taking out an insurance". In terms of accountability, this distinction could well make a difference.

Overall, while the PROGRESA study evaluated the intervention under normal conditions and as it was planned to operate in the future, Al Amana and SKY both evaluated programmes implemented in "non-normal" conditions. This obviously limits the use of the results for summative evaluation purposes.

We now turn to the issue of the varying intensity of the programmes with respect to the different beneficiaries. The term "intensity" is understood to mean the amount of cash transfers (PROGRESA), microcredit (Al Amana) or health insurance (SKY) that the different households could access. In the case of PROGRESA, the level of intensity was built into the design inasmuch as households were entitled to a fixed amount of transfers based on family demographics and could not change that amount. In the case of Al Amana and SKY, however, the households themselves could decide how much of the programme they would "consume", which results in very varying intensities of endogenous treatment. In Al Amana, some households may have borrowed different amounts or may have borrowed several times, while others may have only borrowed only once. In addition, take-up was observed to be very progressive within communities, which means that some households used the credit early on while others had barely started to borrow at the time of the second-round survey. SKY experienced a similar issue in that many of the households dropped out of the insurance scheme over the course of the study once the grace period had terminated. As implemented, however, the study pulls together households that had been insured for two years and households that had only been insured for six months.

In the cases of both Al Amana and SKY, these non-constant treatment intensities clearly limit interpretation of the studies' impact results. Obviously, the analyses could try to differentiate the levels of impact according to the different intensity of borrowing/insurance. However, there are two limitations to this approach: first, it is not possible to run impact estimates on subsamples of the dataset, and secondly the

experiments fail to take into account the actual endogeneity of the intensity.

A final issue relates to the novelty of the programmes. As mentioned earlier, PROGRESA was quickly taken up by its potential beneficiaries. In contrast, microcredit in the remote rural areas of Morocco was clearly not well understood by a large portion of the population. As a matter of fact, a qualitative study undertaken by Guerin *et al*. (2010) on a subsample of the villages involved clearly showed that households had very varied perceptions of the use of credit and repayment obligations (for instance, the latter varied depending on whether Al Amana was perceived as a government entity, in which case the loans were mostly understood to be transfers). The same is true in the case of SKY, where households had to understand an entirely new concept whereby their health costs were covered by their insurance scheme. In both cases, it is likely that after a few years of experience the reactions of the households will be very different to those during the trial-and-error period of the first few months.

Overall, both the Al Amana and the SKY studies pose significant challenges with respect to their capacities to interpret the results. This again limits the use of the studies for summative purposes.

2.3.3. On what is impact measured?

A final issue relates to the outcome that the programme seeks to achieve and how it can be measured within an impact evaluation. In the case of PROGRESA, clear and easily measurable indicators directly related to the Millennium Development Goals (MDGs) were established regarding school attendance and visits to health centres. Further, the impact on these outcomes was expected in the very short run, since the transfers would stop after three months if the conditions were not fulfilled. In other words, the causal chain was very short. The impact on PROGRESA's other targets such as empowerment of women is more challenging to assess because of measurement problems and the time it may take for impacts to occur. In fact, the PROGRESA IE relied heavily on qualitative studies to assess these latter aspects (Skoufias, 2000).

The outcomes immediately impacted by microcredit (e.g. agricultural production) are not as directly relevant to discussion of the programme's contribution to MGDs as the outcomes of the PROGRESA programme. On the contrary, it is reasonable to expect that the impact of microcredit on poverty-reducing outcomes will take a relatively long time to materialise. In fact, the first loans granted involve small amounts and a number of phases must be completed before the investments financed translate into decreased poverty.[13] In addition, and as discussed above, a learning curve is likely to be associated with the use of such a new product. Overall, it may take a relatively long time for the impact of microcredit to materialise into changes in poverty outcomes. Finding no statistically significant

[13] In fact, results from the Spandana study in Hyderabad suggest that the causal chain between access to microcredit and poverty is long, heterogeneous and complex (see Banerjee *et al*., 2010).

effects in the short run could therefore be misleading.[14]

The issue is different in the case of SKY. Here, the main idea was to measure the impact of health insurance on health-seeking behaviour, health-related outcomes and debt burdens after an accident or illness. The IE reveals that variations in the factors most closely tied to the MDGs (notably maternal and infant health) are very small and the occurrences of severe health-related shocks so rare that it was difficult to find statistically valid variations. And while the study finds impact on intermediary outcomes, the link between these outcomes and the donors' need for MDGs-framed results is not complete.

Thus, in the case of Al Amana and SKY, the summative use of IE results is restricted by the limited time frame of the study and the main outcomes initially targeted (poverty in one case, lower indebtedness in the other), which are both expected to be statistically detectable after a significant length of time. In fact, the types of IE described here are most often poorly adapted to assessing medium-term impacts. The major constraint here is that it is difficult to keep a control group immune from contamination by the programme for too long. And while some attempts have been made to evaluate longer-run effects of PROGRESA-like programmes, they have to rely on variations in the number of months of programme exposure, which significantly affects statistical power.

[14] It has been argued that finding no short-run impact on poverty is itself an indication that programmes with faster effects should be promoted instead. And in fact, it is true that programmes such as social transfers may work faster. This however forgoes the recurring question of how sustainable the impacts themselves prove to be.

2.4. Conditions for summative impact evaluations

What lessons can be drawn from these three examples with respect to the use of IEs for accountability needs? We use the differences in the nature of these three programmes, together with the evaluative questions effectively addressed in their respective IEs, to propose a rule of thumb for identifying the kind of programmes and questions that allow IEs to meet accountability requirements.

We will first extract some key components on which these programmes differ and which are relevant to our discussion on impact evaluations. First, relating to "on whom the impact is measured", the three programmes vary in their beneficiaries' propensity to participate and thus in their take-up rates. In the case of PROGRESA, a large proportion of the beneficiaries were expected to participate and nearly all targeted individuals indeed chose to accept the payment and the conditions attached. In the case of Al Amana and SKY, however, only a fraction of the population was expected to do so. This led to very purposive sampling schemes likely to affect the studies' capacities to generate results that are representative of the underlying targeted population.

Second, relating to "the impact of what is measured", the programmes differ with respect to the novelty and complexity that they present to beneficiaries. Clearly, PROGRESA constituted an innovation for rural households in Mexico but results from the study indicate – *a posteriori* – that its mechanics (and in particular the conditions attached to the transfers) were fairly easily understood by the targeted beneficiaries, and the programme design was thus not changed during the course of the study. In contrast, the qualitative analysis undertaken of the Al Amana and SKY programmes (cf. Guérin *et al.*, 2010; Portejoie *et al.*, 2008; Ramage *et al.*, 2010) revealed that the proposed products were quite new and complex. A slow learning process was thus to be expected on the demand side not only regarding the choice to participate, but also the use that beneficiaries would make of these products. This, in turn, can affect the type of impact that is measured.

For this second point, however, homogeneity of treatment appears to be an important point of difference between the programmes. For PROGRESA, cash transfers are not homogeneous amongst beneficiary households, but the conditions attached to the transfers are the same for all. Homogeneity of treatment could also be seen in terms of the time period, in that every beneficiary household belonging to the same wave of the programme's scaling-up is "treated" by RCT for exactly the same period of time.

In contrast, it is the heterogeneity of treatment that characterises the Al Amana and Sky programmes. We have seen above

that intensities of treatment were variable for the beneficiary households according to the amount and the number of loans subscribed or the duration of insurance cover. Heterogeneity also pertains to the point in time when the treatment is taken up. In both cases, the loans and insurance could be taken up by households at any time between the baseline and endline surveys. Importantly, these differences in intensity depend on the choice of the households themselves, not of the programme. Finally, for Al Amana, the conditions of the treatment and particularly the interest rates are different between households depending on the period considered.

Third, in connection with the question "On what is impact measured?», these programmes differ in the expected length of time necessary for the intervention to reach its objectives. In fact, while PROGRESA, Al Amana, and SKY all target long-term poverty alleviation, short-term intermediary outcomes with a direct link to poverty (for instance through the MDGs) are more or less easily identifiable. The impact of PROGRESA on outcomes such as school attendance or visits to health centres can be observed within just a few months (households would otherwise immediately lose access to the transfer). In contrast, the impact of microcredit on poverty may take longer to become visible, and the relationship with short-term impacts can be rather tricky (see for example Banerjee *et al.*, 2010). In the case of SKY, impact can only be measured if significant health shocks occur, and these can only be statistically detected if quite a large number are observed. Here also, measurable impacts are therefore more likely to be identified in the medium- to long-run.

Table 1 *Selected differences between Progresa, Al Amana and SKY*

	Participation rates	Learning curve involved	Homogeneity of treatment	Length of time for impact to occur
PROGRESA	Close to 100%*	Limited	Strong	Short term
Al Amana	13.6%**	Considerable	Weak	Medium/long term
SKY	27% ***	Considerable	Medium	Medium/long term

**share of eligible families; ** share of purposive sample; ***share of total population*

Table 1 summarises the factors that determine IE's capacity to produce summative assessments and to meet the donor requirement for accountability. In particular, it highlights some of the reasons why the PROGRESA IE satisfied accountability needs whereas the Al Amana and SKY evaluations were less able to do so.

Two points are worth mentioning at this stage. First, most of these constraints essentially apply to IEs that are planned before the programme is actually implemented. In this case, and particularly when an innovative approach is applied, take-up rates are most often unknown *ex ante*, and the programme's design is not yet stabilised. Such constraints are particularly relevant in the case of IEs that affect the intervention itself, either through the choice of beneficiaries/non-beneficiaries, or through their influence on the programme content itself – for instance, on the pricing policy of the service provided. Randomised control trials such as the ones described above, while arguably the most statistically robust, are also the most typical examples of studies that have a limited capacity to provide summative assessment of a programme's impact.

Second, it would be a mistake to extrapolate from this table that the PROGRESA programme is simpler than the two others. PROGRESA is a large and complex programme with several sub-components on both the demand- and supply-side of health and education. What is perhaps simpler, or at least more focused, is the main policy question raised by the programme. The main concern of the PROGRESA IE was to evaluate the effects of the conditions linked to the cash transfers. In this sense, it is the evaluative question, and not the programme *per se*, that appears more manageable in the case of PROGRESA than in the cases of Al Amana and SKY.

Overall, the above discussion suggests that only a subset of development programmes is suitable for summative IE. The conditions conducive to a summative use of RCTs include: an observation period in coherence with the logical chain; a limited number of well-specified homogeneous inputs; a tried-and-tested process; a short and external-event-proof causal chain; a rapid and stable take-up rate; a high and stable level of participation; a set of outcomes that are measurable in the short run; and/or impacts covering the main aspects of the programme or, at least, the main aspects of the evaluative questions. A useful analogy to describe such programmes is that of a tunnel with a clearly delimited beginning and end, where one can easily define what enters as an input and what is expected to exit as an output, where mid-way drop-out is nigh on impossible, and finally where the path is both short and predictable in that it is immune to external influences. Unsurprisingly, the characteristics of tunnel programmes match nicely with those of medical experiments, from which the IE movement drew its inspiration.

2.5. Conclusion

It is widely recognised that not every programme is adapted to IE, and particularly to RCT methodology. For instance, such methods only apply to programmes where a large number of treatment units can be compared to a large number of control units. IEs are thus suited to micro-level types of interventions, such as those where beneficiaries are individuals, households, classrooms or local communities. This means that not all development programmes are suited to IE.

Our analysis has used three concrete examples to further delimit the characterisation of those programmes that are adapted to RCTs aimed at producing evidence for direct accountability purposes. We suggest describing these programmes as "tunnel" programmes insofar as they satisfy the following requirements: (i) a period of observation coherent with the logical chain, (ii) a limited number of well-specified homogeneous inputs, (iii) a tried-and-tested process, (iv) a short and external-event-proof causal chain, (v) a rapid and stable take-up rate, (vi) a large and stable participation, and (vii) a set of measurable outcomes in the short run, covering the main aspects of the programme.

Accountability is an important aspect of evaluation and there are high expectations, particularly from donors, that impact evaluation could play a more decisive role in this field. However, on this count, AFD's experience could be considered as evidence of the limits of IE's potential to enhance accountability, as well as an invitation to temper these expectations.

The purpose of our article is not to deny that accountability is an objective targeted by IE. Nevertheless, there are many other objectives, including: to provide robust and cumulative evidence on development policies (on pricing, access, etc.); to help provide the missing links between academic research and development practices; to promote trial-and-error processes in policy-building, by conducting experimental programmes prior to scaling-up; to have more influence on policymakers; to test theories and to learn about household behaviour patterns. The Al Amana and SKY evaluations, which this article examines with respect to their limitations in providing summative assessments of a programme's impact, have nonetheless come up with some interesting results. These include, for instance, the first rigorous analysis of adverse selection in poor countries and a precise mapping of the very contrasted loan take-up rates in different areas of rural Morocco.

IE specialists have generally channelled the use of such methods towards field experiments intended to inform the design of future policies, rather than evaluations focused on existing policies. (Duflo, 2009). For donors, however, the use of IE results for accountability purposes is still an important motivation. It is this latter point that this paper has attempted to qualify.

References

Banerjee, A. (2007), "Making Aid Work", in A. Banerjee, *Making Aid Work,* MIT Press, Cambridge MA.

Banerjee, A.V., E. Duflo, R. Glennerster and C. Kinnan (2010, June), "The miracle of microfinance? Evidence from a randomized evaluation".

Banerjee, A. and E. Duflo (2009), "The Experimental Approach in development Economics". *Annual Review of Economics*, pp.151–178.

Barham, T. (2005), "Providing a Healthier Start to Life: the Impact of Conditional Cash Transfers on Infant Mortality", Working Paper, University of California, Berkeley.

Center for Global Development (CGD) (2006), *When Will We Ever Learn? Improving Lives Through Impact Evaluations,* Center for Global Development, Washington D.C.

Crépon, B. and E. Duflo, (2006), "Projet de recherche : Evaluation de l'impact d'un programme de micro-crédit en milieu rural, AFD", Paris.

Crépon, B., E. Duflo and W. Pariente (2007), "Rapport étude de faisabilité : Evaluation de l'impact d'un programme de micro-crédit en milieu rural", AFD, Paris.

Crépon, B., F. De Voto, E. Duflo and W. Pariente (2010*a*), "Evaluation de l'impact du micro-crédit en milieu rural au Maroc", Interim report, AFD, Paris.

Crépon, B., F. De Voto, E. Duflo and W. Pariente (2010*b*), "Evaluation de l'impact du micro-crédit en milieu rural au Maroc", Evaluation report, AFD, Paris.

Deaton, A.S. (2007), New Democracy Forum, in A. Banerjee (ed.), *Making Aid Work* (pp. 55–62), MIT Press, Boston.

Duflo, E. and M. Kremer (2003), "Use of Randomization in the Evaluation of Development Effectiveness", *World Bank Operations Evaluation Department (OED) Conference on Evaluation and Development Effectiveness*, World Bank, Washington, DC.

Easterly, W. (2003), Can Foreign Aid Buy growth? *Journal of Economic Perspective* , 17(3): 23–48.

Fizbein, A. and N. Schady (2009), *Conditional Cash Transfers, Reducing Present and Future Poverty*, The World Bank, Washington DC.

Gertler, P.J., S. Martinez, P. Premand, L.B. Rawlings and C.M.J. Vermeersch (2010), *Impact Evaluations in Practice*, Word Bank, Washington D.C.

Guérin, I., S. Morvan-Roux, M. Roesch, J.Y. Moisseron and P. Ould Ahmed (2010), "Analyse des déterminants de la demande de services financiers dans le Maroc rural", AFD, Paris.

Jones N, H. Jones, L. Steer and A. Datta (2009), Improving impact evaluation production and use, ODI Working Paper 300, ODI, London.

Kanbur, R. (2005), "Goldilocks development economics (not too theoretical, not too empirical, but watch out for the bears!)", *Economic and Political Weekly* , 40(40).

Karlan, D. and J. Zinman (2010, January), Expanding Microenterprise Credit Access: Randomized Supply Decisions to Estimate the Impacts in Manilla, *Yale working paper*.

Khander, S.R., G.B. Koolwal and H.A. Samad (2010), *Handbook on Impact Evaluations: Quantitative Methods and Practices*, World Bank, Washington D.C.

Levine, D. (2010), "Assessing the Effects of Health Insurance: the SKY Micro-Insurance Program in Rural Cambodia", Impact Analysis No.4, AFD, Paris.

OECD (2005), *Paris Declaration and Accra Agenda for Action*, accessed March 2011, at OECD: http://www.oecd.org/document/18/0,3343,en_2649_3236398_35401554_1_1_1_1,00.html

OECD/DAC (2002), Glossary of Key Terms in Evaluation and Results Based Management, OECD-DAC, Paris.

Portejoie, C. and M. Goursat (2008), Member satisfaction survey, GRET, SKY, Phnom Penh.

Pritchett, L. (2002), "It pays to be ignorant: a simple political economy of rigorous program evaluation", *Journal of Policy reform* , 5(4): 251–269.

Ramage, I., K.H. Ramage, E. Mazard, M. Kavenagh, G. Pictet and D. Levine (2010), Village monographs, Summary Report, DOMREI, Phnom Penh.

Ravallion, M. (2008), *Evaluation in the practice of development,* World Bank, Washington, D.C.

Skoufias, E. (2000), *Is Progresa Working? Summary of the Results of an Evaluation by IFPRI.* International Food Policy Research Institute, Washington D.C.

Conférences & Séminaires

Part 3: Is Indicator-Based Management a Guarantee of Efficiency?

3. History revisited: Measurement for Management in Development

Jodi Nelson, Bill & Melinda Gates Foundation

Abstract

This paper considers the optimal use of indicator-based management in development. Situated within the conference's focus on evaluation and its "discontents", the premise is that the development community's failure to learn is not a failure of evaluation or measurement more broadly, but instead a failure of strategic clarity. I provide a brief and necessarily cursory analysis of the logical framework approach, the experience of results-based management in development, relevant critiques of the Millennium Development Goals and a recent book about measuring performance in business. All of these stories point to the same lesson – that indicators are only helpful measurement tools if they reflect an underlying strategy to produce development results. Indicators that are "strategy independent" – in David Apgar's words – are irrelevant. I suggest that the real challenge to the development community's ability to learn from practice does not lie in measurement per se *(in this case, the use of indicator tracking), but in the rarity of our strategic clarity.*

The AFD-EUDN conference asks us to reflect on a perennial question in development: why can't we learn from our own experience? The question is not rhetorical. It can be heard in the frustration of practitioners everywhere who lament the "reinventing of the wheel" that happens in their headquarters and field operations around the world. Given its intention as a tool for decision-making, learning and improvement, evaluation's place in the hot seat at the 2012 AFD-EUDN conference makes sense. The conference organisers ask: with the many different approaches in the evaluator's toolkit, how is it that we seem unable to translate experience into practice to implement better, experience-based development strategies? Is it the way evaluations are designed and done, or the context in which development occurs that "severely reduces the usefulness of past experiences for designing future projects"?[15]

But is the development community's failure to learn really a failure of evaluation? I argue that one of the reasons that evaluation – and measurement more broadly – does not help us learn from experience is that we have failed to learn from our own experience with it. The last few decades are rich with lessons that should be consolidated into conventional wisdom by now. Our failure to make them explicit and build on them impedes our ability to use measurement as the powerful tool it can be in the development community's many efforts to achieve results that matter for people.

One of the most important lessons is relevant to the question I was asked to consider as part of the conference: *What is the optimal use of indicator-based management to evaluate development results?* I propose that the answer to this question is clear from development experience both at the field level, where development programmes are designed, and at the policy level, where Western donor agencies have sought to institutionalise results-based management (RBM) in recent years. The lesson is that measurement is only as good as we are. In other words, if we do not know where we are going – the results we hope to achieve for people and how to produce them – any road will do and measurement is useless as a consequence. Recognising this lesson helps us to see that the optimal use for indicator-based management is precisely what the term suggests – to measure the progress of a particular strategy and use the resulting data to manage its implementation. This puts into question the utility of high-level indicators such as those associated with the Millennium Development Goals (MDGs) because they are disconnected from evidence-based strategies for how donors, governments, implementing agencies and ultimately communities can reach them.

Although I will reference and use the MDGs to make my argument, the paper does not comment on the measurement issues directly related to the MDGs since there is a wide range of critique and discussion about this topic in the literature.

[15] Cf. presentation of the 2012 AFD-EUDN conference at http://www.afd.fr/lang/en/home/presse-afd/evenements/conference-eudn/EUDN2012.

3.1. Measurement is only as good as we are

Indicator-based management has definitely been embraced by the development field. But this is not a new trend. The MDGs and aid efficiency indicators of the Paris Declaration are only the most recent and currently visible high-level set of results, indicators and related targets. In fact, indicators have been prominent tools in development work for decades – both in the everyday field operations of development organisations and in the headquarters of the largest donor agencies. These experiences suggest the value of strategic clarity for good measurement.

3.1.1. What does the logical framework approach tell us?

Indicators have always been a key component in the nuts and bolts of development programming – an enterprise in which the basic approach suggests a lesson that is manifest in the many manuals, policies and training actions designed to equip practitioners with the tools of programme design and programme cycle management. The lesson in brief is that indicators can only be useful measurement tools if it is clear what you are trying to do and how. This sounds intuitive, but most evaluators have had the experience of being asked to evaluate a programme whose results and logic are not clear.

It is no surprise that evaluators were in part responsible for creating the toolkit that development practitioners use to design their programmes today. When evaluation gathered momentum in development in the 1970s in USAID and in a few of the larger UN agencies, project objectives were not clearly specified, nor was it clear what the related work was intended to achieve. Methods for programme planning were disorganised at best, making the old adage "if you don't know where you're going, any road will take you there" especially relevant.

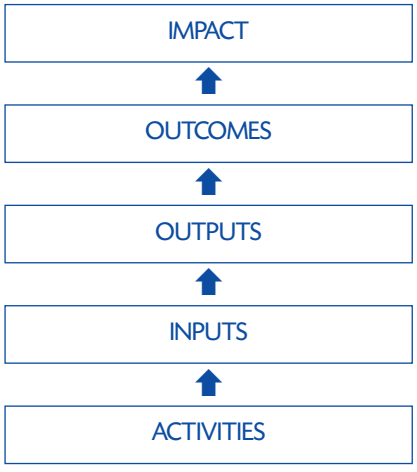
In his survey of the history of development evaluation, Basil Edward Cracknell describes how "trying to evaluate in these circumstances was like trying to ride a bicycle with loose handlebars" (Cracknell, 2000). He explains how the evaluation staff in the bilateral development agencies promoted results-driven programme design as an essential starting point for both implementation and measurement. They introduced the logical framework tool that remains a centrepiece of the programme cycle management approach used by many if not all mainstream organisations today.

Although practitioners lament its varying donor formats, the template itself is less important than the essence of strategic thinking that is intended to drive the planning process and help practitioners to define measureable goals, a path to achieve them and a set of relevant indicators. The "logic" of the framework refers to the causal relationship between inputs, activities,

outputs, outcomes, impact – the ultimate or highest level of change sought for people. The implicit rationale of the approach is that if certain activities are undertaken and inputs provided, given certain assumptions, a set of outputs will result; and this set of outputs, again given certain assumptions, will lead to outcomes; and this set of outcomes, again given certain assumptions, will lead to impact. This "means-ends relationship," or the vertical logic, is displayed pictorially below.

Figure 1

The vertical logic framework



This essential deductive reasoning is probably the most important component of good measurement. Given early ideas that aid was more about charity and good intentions than a means to strategic ends (Smillie and Minear, 2004), it makes sense that the turn to strategy came rather late to development. Even as recently as the late 1990s, people were still commenting on the value of what was seen as a new approach to getting work done. As George Foulkes, Parliamentary Under-Secretary of State at DFID said in 1997:

> The regular use of logframe matrices with their requirement for specific indicators and means of verification, has not just brought clarity to project design, it has extended our capacity for monitoring and evaluation. We now have to define in advance not just what we are seeking to do, but how we shall know whether, and when, we have achieved our goal. (in Cracknell, 2000)

3.1.2. What does the experience with Results-Based Management tell us?

Economic crises and public sector reform across the OECD countries in the 1990s reinforced the need to take results-driven strategies from project to institutional levels. USAID, the World Bank, UNDP, UNICEF, AusAID, DFID, NORAD and others integrated results-based management approaches into their operations, seeking to replicate private sector tools and incentives to assure accountability for results in the public provision of services. Strategic planning, measureable goals, precise targets, performance monitoring plans and rewards for staff on the basis of performance show up in spades across the policy statements and guidelines of all the major Western donors at the time.

The case for the change was compelling. Popular catchphrases such as "demonstrating value for money" and "doing more with less" foreshadowed trends today – when changes in political and economic circumstances have predictable implications for the way we

conceive of development and its measurement. In a time of scarce resources: "Simply measuring success by the volume of spending, or even the number of teachers trained, kilometres of road built and women's groups formed, is a not a satisfactory approach. Input monitoring does not ensure that development spending makes a difference to people's lives". The rise of results-based management shifted their focus from project-level planning to overall performance of country programmes and their organisations overall (White and Black, 2004; Hulme, 2007).

A few statements from the relevant agencies show just how much donors aligned on a similar approach.

- At the World Bank: "Results based management provides a coherent framework for strategic planning and management based on learning and accountability... It is first a management system and second, a performance reporting system" (WB, 1997).
- At the Canadian International Development Agency (CIDA): "Introducing a results-oriented approach . . . aims at improving management effectiveness and accountability by defining realistic expected results, monitoring progress toward the achievement of expected results, integrating lessons learned into management decisions and reporting on performance" (CIDA, 2009).
- At USAID: The Office of Management and Budget designates USAID's sustainable development activities as a pilot project for performance measurement for FYs 1995 and 1996 under the Government Performance and Results Act. As a pilot, USAID is committed to expanding and deepening strategic management in more than 40 sustainable development programmes, better linking performance measures to Agency programming and management systems, and testing broader management reforms aimed at enhancing the Agency's ability to manage for results (Britan, 1998).
- At UNDP: The objective of RBM is:

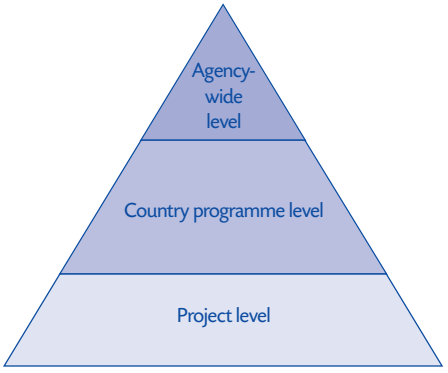
  "to provide a coherent framework for strategic planning and management based on learning and accountability in a decentralised environment.' Introducing a results-based approach aims to improve management effectiveness and accountability by 'defining realistic expected results, monitoring progress toward the achievement of expected results, integrating lessons learned into management decisions and reporting on performance." (UNDP, undated, p.3)

Agencies acknowledged a similar set of requirements for the approach to work. Most relevant for this discussion is that they took as a starting point measureable goals and strategic plans to accomplish them, seeking to connect long-term ends at a high level with the day-to-day activities of programme managers and staff. The idea was that this type of analytic thinking and strategising would help align the results of different management levels of their operations – from project, to country or operating units, to the corporate or agency-wide level.

Independent reviews by the US Government Accountability Organization (GAO) and the Auditor General of DFID confirm that the

Figure 2

Results-based management at different organisational levels



jump in logic from project to organisation was the most challenging to accomplish because of a lack of strategy.

> USAID has made progress in establishing outcome-oriented goals and developing indicators and targets that help measure overall results. However, because the agency's goals in the three outcome areas are so broad and progress is affected by many factors other than USAID programs, the indicators cannot realistically serve as measures of the agency's specific efforts. USAID recognizes this limitation and has improved its fiscal year 2001 performance plan to discuss agency efforts within this broader context. USAID is also seeking to better understand the relationships between its specific programs and their contributions to the desired overall outcomes. (GAO, 2000: 1-2)

GAO notes further that USAID's programme strategies constitute only a small part of the overall strategy for achieving progress and that it lacks a strategy for how these specific programmes relate to broader goals: "USAID seeks to reduce the proportion of the population in poverty, but its major program efforts in this area emphasize the development of microenterprises and nonfarm rural enterprises. The agency does not clarify how closely linked these specific programs are to the overall goal" (ibid). GAO suggests that USAID should use proxy indicators that are closely related to the work it actually does on the ground.

DFID's experience was similar. "The key design challenge", states the Auditor General in a 2002 National Audit Office (NAO) report, "is the extent to which performance measures adequately reflect DFID's contribution to outcomes. The difficulty of establishing firm links between DFID's work and the achievement of outcome-oriented development goals is faced by all development agencies". The auditor cites an NAO-commissioned paper by Howard White which states that it is "impossible (or at best virtually impossible) for an individual agency to isolate its impact on global or even country trends in the [International Development Target] indicators" (NAO, 2002).

Finally, UN reviews found the same weakness in strategy to be a key obstacle to useful performance measurement. For example, the 2004 Joint Inspection Unit report on UN performance related to achieving the universal education goal (Education for All) found that four years after adopting the related goals, there had still not been

concerted effort to plan their implementation. The JIU report noted that UN objectives were not "logically consistent among levels" and that "making results fit together and add up into major outcomes for the organization as a whole is what designing strategies is all about". The report concludes that "[i]n all cases, it is vital that the organization works to avoid a strategic disconnect in its programming at the various levels if it is to implement RBM successfully" (UNJIU, 2004).

3.1.3. MDGs in context – Whither the strategy?

As articulated by the OECD and the UN soon thereafter, the International Development Targets and then the Millennium Development Goals were part of the same new public management that drove donor agencies to institutionalise results-based management.

> The goals are set in precise terms– measured in numbers to ensure accountability. The openness and transparency of such numbers can help us chart a course to achieve the goals and track progress. These goals are worthwhile because they will improve the quality of human life. The world will be better, and safer, for its 6 billion people and for the projected 7 billion people in 2015. (IMF, OECD, UN and World Bank Group, 2000)

Although the MDGs satisfy several criteria for being good performance measures, their lack of connection to strategy is problematic and noted by several commentators. In White's words:

> they are outcome oriented with little effort made to build a consensus around an underlying logic model of how the targets are to be achieved. In the absence of such a model, it is extremely difficult to say anything sensible in terms of attributing change in target indicators to the actions of the development community. Targets do not in themselves contain the strategy as to how they should be attained. (White, 2004)

Luik offers a harsher critique:

> As the most basic of planning manuals notes, goals without carefully crafted and detailed plans are useless since they offer up a destination but neglect the roadmap necessary for finding it. In fact they are worse than useless since they give the appearance that something, as opposed to nothing is happening. But the UN's MDGs, like some disappointing children's toy, come not only without batteries but without any guidebook for assembly. They are a vision ungrounded in any sense of the strategic steps necessary to bring the vision to fruition...disconnected from a credible strategic plan as to how such difficult goals can be met. (Luik, 2005)

If the MDGs are not good tools to measure development results because they are not linked to an overall strategy to achieve them, what use are they? Many analysts and even policy makers engaged in their creation note that the MDGs were neither designed nor intended to be planning targets. Kofi Annan's 2005 report on the progress of the Millennium Development Declaration described the MDGs as "globally accepted benchmarks of broader progress". Sakiko

Fukuda-Parr – a principal author of many of the Human Development Reports – characterises the MDGs as "political commitments, made by world leaders, that define priorities in a normative framework and that can be used as benchmarks in evaluating progress" (Fukuda-Parr and Greenstein, 2010). Indeed if the MDGs were meant to raise awareness of neglected global issues, build consensus among world leaders, mobilise attention and induce them to pledge to take concrete actions, it is possible that their purpose has already been achieved. As the authors put it, "Like all UN Goals, the MDGs are best viewed as a set of norms... They are ends rather than means, and do not come with a specific set of new development strategies for meeting the goals" (ibid).

Michael Clemens and his co-authors Charles Kenny and Todd Moss agree with this, arguing that the MDGs can be understood in two ways: as real targets of the development community – a basis then for measuring progress and evaluating results – or as a symbol of the kinds of outcomes toward which the development community should strive. He suggests they are the latter: "useful benchmarks that publicly bring out the stark contrast between the world we want and the world we have and cause us to redouble our search for points of intervention to close the gap" (Clemens *et al*., 2007:).

Jan Vandemoortele, co-architect of the MDGs, supports this view, suggesting that global targets only apply at the global level and that it is not appropriate for the MDGs to be used as yardsticks for measuring and judging performance at the national level. He characterises this approach – using the MDG indicators as the basis for performance measurement – as suffering from misplaced concreteness: "Their interpretation as one-size-fits-all targets abstracts away the specific and historical background of each country, its political system, its natural endowment, its geography, its internal divisions, and other challenges it may face. The post-2015 targets must guard against the misconception that global and national targets are one and the same" (Vandemoortele, 2009: 5).

Because they are global and not national or local targets, assessing whether progress is "on track" for meeting the targets can only be done at the global level. If you accept this rationale, concludes Vandemoortele, it is wrong to "lament that Sub-Saharan Africa will not meet the MDGs. These targets were not set specifically for that region" (Vandemoortele, 2007).

3.2. Putting the pieces together – the centrality of strategic clarity for indicator-based management

3.2.1. Strategic clarity in business

We often think of the private sector as better poised to measure its performance because of the natural feedback loop that comes with the market. In other words, you know exactly what you are looking for – money and market share – and can gather data relatively easily on both because these are observable, measureable phenomena and the data are readily available. But the importance of strategic clarity for measurement is central to private sector experience as well.

In his book, *Relevance: Hitting Your Goals by Knowing What Matters*, David Apgar reflects on the challenge facing firms because of their declining ability to learn from experience as they accumulate more and more data of limited relevance. He describes how businesses can articulate and measure requirements for meeting a financial goal rather than devising a specific strategy to meet it: "just as you can know all the ingredients for a dish without having a recipe for it, you can meet every conceivable requirement for a goal without knowing how to achieve it. Simply put, requirements are not strategies" (Apgar, 2008:). Commenting on the popular balanced scorecard approach used to define and measure performance by firms, Apgar shares survey data that suggest how unhappy executives are with this approach and how little evidence there exists that significant improvements in returns on sales or assets result for those firms that use it.

Using cases across recent business experience, Apgar argues that "knowing what matters demands a planning and performance approach that derives performance metrics or indicators from key strategic assumptions – not from balanced lists of outputs and inputs they require" (ibid). He calls these measures – the ones derived from strategy – "relevant" as they are the ones most relevant to what firms are trying to do. Balanced scorecards provide a performance measurement framework that adds non-financial performance measures to traditional financial metrics to give executives a more "balanced" view of organisational performance (emphasising indicators on customer satisfaction, efficiency of internal process, learning and development) but its measures are "irrelevant to the question of whether a particular strategy is succeeding as intended" (ibid). With examples of prominent firms that have had the same scorecard even though their strategies have changed many times, he notes how irrelevant it is to measure indicators on a scorecard that is "strategy independent" (Apgar, 2008).

Case studies of BP, Alcoa and GE help Apgar to make a strong case that the same "strategic clarity" that governments and development organisations need to measure their results and learn from experience is needed for business to do the same. For him, performance strategies need to be explicit enough to be testable, with short- and medium-term progress indicators that are logically connected to how a firm will achieve its goals.

3.2.2. Strategy in development – whither the theory?

Our experience with "doing" development contains enough lessons to confirm that strategy – whether at the project, programme, country or organisational level – is as important to successful measurement of results as it is in the private sector. It is embedded in the very tools that practitioners use around the world to plan for and implement their work. It is also evident in the experience donor organisations have had trying to institutionalise results-based approaches that link project to programme to country and organisation without sufficient attention to logic to connect the three. Going back to my initial question of whether or not the development community's failure to learn from experience is genuinely a failure of evaluation, I believe it is more accurately a failure of learning from experience that strategy, based on evidence of what works to achieve results for people, is the most essential missing piece to the cycle between planning for, measuring and learning from results to improve development practice.

In the late 1990s, Hugo Slim described the "crisis of theory" that existed to support the then new move to connect relief, development and peace in the aid enterprise. Although there was a general consensus that all three components should make up an appropriate path to help countries recover from war, and donors and NGOs had a sense of how important these were, they had very few theories and even less evidence to help get there. While development as a proper field of discipline is much stronger than ever before, there exists the same crisis of theory, and therefore of strategy to accomplish results.

The lack of theory at a micro level – where change happens in a country or community – is where we should be focussing more attention, rather than tracking macro-level indicators that are disconnected from strategies to help improve people's lives. Concrete, evidence-based strategy can help define more useful indicators for management and clarify accountability for achieving results. In a response to Hilary Benn's articulation of DFID's 2006 government White Paper on poverty alleviation, William Easterly frames the issue well, noting how lofty objectives that are breath-taking in scope do not move us any closer to accomplishing them. As Easterly writes: "Surely, whether peace, prosperity and democracy break out depends on a few other things besides what DFID does. Exactly how meaningful is a promise to achieve things so far beyond your control? How could anyone hold you to account for whether such promises are kept?" (Easterly, 2006).

Even Jeffrey Sachs, the global cheerleader for the MDGs, acknowledges the lack of strategy underpinning the high level MDGs. He is

quoted in Neil MacFarquhar's *New York Times* article saying: "There is no plan of action to complement what will be agreed upon... There is a difference between lurching forward on good intentions and a range of unconnected initiatives versus having a high priority plan" (MacFarquhar, 2010). Esther Duflo is quoted in the same article: "The goals serve as a useful wish list for what everyone on the planet should have access to for a decent life, but the glaring hole is the absence of information about why they work in some places and not in others" (ibid.). She goes on to point out that, although there should be fewer children dying in northern Nigeria thanks to the available health services, these do not seem to be liked or used by the local people: "We don't understand what works. We all made the goals and said, 'Let us get there,' without asking what we really know about how to get there" (ibid.).

Putting all these pieces together, what can we say about the optimal use of indicator-based management as an approach to measure the results of development?

1. Indicator-based management is best used when the indicators themselves are derived from a clear and logical strategy for accomplishing them. Back to Apgar, the indicators cannot be "strategy independent" if they are to help us learn from experience. Moreover, management requires being able to adjust things that are within an organisation's control to change or at least close enough that a change in practice will be reflected in observable changes in the indicator tracked. My read of the history of development is that this lesson is and has been clear in the business of actually doing development work – at the micro level, where organisations plan and design programmes to deliver goods and services, partner with government to improve systems and change people's lives. Here, the data collected to measure progress can genuinely be used to improve management and the work itself.

2. The farther away in terms of in logic and space indicators get from the work itself, the less useful they are because they are not measures of progress of an intentional plan or strategy. I think this lesson has already been learned through experience, trying and failing to connect results at project, country and organisational levels. Although the policy need to make the case for alignment is compelling, it is not clear that this has ever been done successfully. Hopeful commentaries on the next phase of MDG planning reflect on the need for more context-specific, national targets linked to resources and constraints at the country level. This level of analysis and planning would reflect learning from experience, as organisations need different sources of information, for different purposes, depending on what they do. Government, implementing agencies, public and private donors have different incentives and requirements for decision-making. The biggest mistake we can make moving forward is not seeing this and instead continuing to assume that every organisation can and should plan, implement and measure toward the same results.

3. We do ourselves a disservice by planning to achieve results and then measuring them when we do not have a hypothesis for how to achieve them in the first place. This core lesson is evident in development and in the private sector experience of performance measurement. Because the relationship between strategy and measurement is essential, the real gap is in our efforts to build up strong strategies at the micro level of development work – where people's needs, community resources, national institutions, policy and history provide the best opportunity to learn from experience and improve development practice.

In conclusion, I look forward to the conversation in Paris about "evaluation and its discontents." I would reframe the title itself, proposing that it is not that evaluation is not doing development right, but that we are not setting it up for success by clarifying, investigating and continuing to adapt strategy that accomplishes results for people on the ground.

References

Annan, K. (2005), "In larger freedom: towards development, security and human rights for all", Report of the Secretary General, 59th session of the UN General Assembly, available at http://www.un.org/largerfreedom/contents.htm.

Apgar, D. (2008), *Relevance: Hitting Your Goals by Knowing What Matters*, Josey-Bass, San Francisco.

Attaran, A. (2005), "An Immeasurable Crisis? A Criticism of the Millennium Development Goals and Why They Cannot Be Measured", *PLoS Medicine*, 2(10), Public Library of Science, San Francisco.

Binnendijk, A. (2000), "Results Based Management in the Development Cooperation Agencies: A Review of Experience", DAC Working Party on Aid Evaluation Background Report, Organisation for Economic Co-operation and Development, Paris.

Black, R. and H. White (eds) (2004), *Targeting Development: Critical Perspectives on the Millennium Development Goals*, Routledge, London.

Bourguignon, F., A. Benassy-Quere, S. Dercon, A. Estache, J. Willem Gunning, R. Kanbur, S. Klasen, S. Maxwell, J.-P. Platteau, and A. Spadaro (2008), "Millennium Development Goals at Midpoint: Where do We Stand and Where do We Need to Go?", European Report on Development 2008, European Commission, Brussels.

Bourn, J. (2002), "Performance Management – Helping to Reduce World Poverty", Report by the Comptroller and Auditor General, Department for International Development, London.

Britan, G. (1998), "A View from USAID [On Performance Measurement, Evaluation, and Results-Based Management]", in *Public Sector Performance – The Critical Role of Evaluation*, World Bank Evaluation Department, Washington D.C.

Canadian International Development Agency (CIDA), "Results-based Management in Canadian International Development Agency: A How-to-Guide", 2009. http://www.acdi-cida.gc.ca/acdi-cida/ACDI-CIDA.nsf/eng/NAT-92213444-N2H

Clemens, M.A., C.J. Kenny and T.J. Moss. (2007), "The Trouble with the MDGs: Confronting Expectations of Aid and Development Success", *World Development*, 35(5).

Cracknell, B.E. (2000), *Evaluating Development Aid: Issues, Problems and Solutions*, Sage Publications, New Delhi.

Easterly, W. (2006), "Is Foreign Aid Working", Letter to Hilary Benn, *Prospect,* Issue 128, 19 November 2006, available at:
http://www.prospectmagazine.co.uk/2006/11/isforeignaidworking/

Easterly, W. (2009), "The Tragedy of the Millennium Development Goals", available at:
http://aidwatchers.com/2009/07/the-tragedy-of-the-millennium-development-goals/

Fukuda-Parr, S. (2004), "Millennium Development Goals: Why they Matter," Global Governance, Vol. 10, The Academic Council on the United Nations System, Ontario.

Fukuda-Parr, S. (2008), "Are the MDGs Priority in Development Strategies and Aid Programmes? Only a Few Are!", International Poverty Centre Working Paper 48, United Nations Development Programme, Brazil.

Fukuda-Parr, S. and J. Greenstein. (2010), "How Should MDG Implementation be Measured: Faster Progress or Meeting Targets?" International Policy Centre for Inclusive Growth Working Paper 63, United Nations Development Programme, Brazil.

Grimm, S., J. Humphrey, E. Lundsgaarde and S.-L. John de Sousa (2009), "European Development Cooperation to 2020: Challenges by New Actors in International Development", European Development Cooperation, Vol. 4, 7th Framework Programme, Bonn.

Gwatkin, D.R. (2005), "How Much Would the Poor People Gain from Faster Progress Toward the Millennium Development Goals for Health?", *The Lancet*, 365: 813–817.

Hulme, D. (2007), "The Making of the Millennium Development Goals: Human Development Meets Results-Based Management in an Imperfect World", Brooks World Poverty Institute, Working Paper 16, University of Manchester Press, Manchester.

IMF, OECD, UN, WORLD BANK GROUP (2000), *2000, A Better World for All: Progress Towards the International Development Goals* (2000), Washington DC., available at:
http://www.imf.org/external/pubs/ft/jointpub/world/2000/eng/bwae.pdf

Jolly, R. (2005), "The UN and Development Thinking Practice," *Forum for Development Studies*, 32(1), Norwegian Institute of International Affairs, Oslo.

Kaplan, R. and D. Norton (1992), "The Balanced Scorecard – Measures that Drive Performance", *Harvard Business Review*, 83(1).

Kenny, C. and A. Sumner. (2011), "More Money or More Development: What Have the MDGs Achieved?", The Center for Global Development, Working Paper 278, Center for Global Development, Washington D.C.

Kusek, J., R. Rist and E. White (2005), "How Will We Know the Millennium Development Goal Results When We See Them?: Building a Results-Based Monitoring and Evaluation System to Give Us the Answers", *Evaluation*, 11(1).

Leroy, M. (December 2011), "'Value for Money' or 'Results Obsession Disorder'?", The Broker Online, Foundation for International Development Publications, Leiden.

Luik, J. (2005), "Millennium Development Holes", *Nature*, Vol. 446, 347.

MacFarquhar, N. (2010), "U.N. Poverty Goals Face Accountability Questions", *The New York Times*, 18 September 2010, available at:
http://www.nytimes.com/2010/09/19/world/19nations.html

National Audit Office (NAO) (2002), "Performance Management – Helping to Reduce World Poverty", Report by the Comptroller and Auditor General, HC 793 Session 2001-2002: 12 April 2002, The Stationery Office, London.

Smillie, I. and L. Minear. (2004), *The Charity of Nations: Humanitarian Action in a Calculating World*, Kumarian Press, Sterling.

Tabatabai, H. (2007), "MDG Targets: Misunderstood or Misconceived?" *IPC-IG* One Pager 33, International Policy Center for Inclusive Growth, Brazil.

UNDP (2002), "Results-based Management: Concepts and Methodology", available at:
http://www.undp.org/evaluation/documents/RBMConceptsMethodgyjuly2002.pdf

UNESCO (2001), "Results-based Programming, Management, and Monitoring Guide",
http://www.unesco.org/bsp/eng/rbm.pdf

UNITED NATIONS JOINT INSPECTION UNIT (UNJIU) (1999), "Results Based Budgeting: The Experience of United Nations System Organizations", Joint Inspection Unit, Geneva, available at http://www.unjiu.org/data/reports/1999/en99_03.pdf

UNITED NATIONS JOINT INSPECTION UNIT (UNJIU) (2004), "Implementation of Results-Based Management in the United Nations Oranizations", Part I, Series on Managing for Results in the United Nations System, Joint Inspection Unit, Geneva.

US GOVERNMENT ACCOUNTABILITY OFFICE (GAO) (1998), "The Results Act: An Evaluator's Guide to Assessing Agency Annual Performance Plans." (1998), *GAO Agency Fiscal Year Performance Plans*, GAO/GCD -10.1.20, United States Government Accountability Office, Washington D.C.

US GOVERNMENT ACCOUNTABILITY OFFICE (GAO) (1999), "Managing for Results: Opportunities for Continued Improvements in Agencies' Performance Plans", *GAO Agency Fiscal Year Performance Plans*, GAO, Washington D.C.

US GOVERNMENT ACCOUNTABILITY OFFICE (GAO) (2000), "Observations on the U.S. Agency for International Development's Fiscal Year 1999 Performance Report and Fiscal Years 2000 and 2001 Performance Plans", *GAO Agency Fiscal Year Performance Plans*, United States Government Accountability Office, Washington D.C.

Vandemoortele, J. (2007), "MDGs: Misunderstood Targets?" *IPC-IG* One Pager 28, UNDP International Policy Centre, Brazil.

Vandemoortele, J. (2009), "Taking the MDGs Beyond 2015: Hasten Slowly", paper commissioned for a High-Level Policy Forum on After 2015: Promoting Pro-Poor Policy after the MDGs, organised by DSA/EADI/ Action Aid; Brussels, June 2009.

Watkins, K. (2011), "The Millennium Development Goals: Three Proposals for Renewing the Vision and Reshaping the Future", available at:
http://www.scribd.com/doc/2442520/Millennium-Development-Goals

White, H. (2004), "Using development goals and targets for donor agency performance measurement", in Black, R. and H. White (eds), T*argeting Development: Critical Perspectives on the Millennium Development Goals*, Routledge, London.

4. How much is Enough? Does Indicator-Based Management Guarantee Effectiveness?

Catherine Paradeise, University Paris Est, LATTS, IFRIS

Abstract

Indicator-based policy steering is staunchly embedded in an underlying technical approach and a management philosophy that draws on a simplified vision of the organisational relationships between policy designers, users, advisors and overseers. This governance model grounded in the principle of accountability raises two questions. First, as regards its efficiency – its capacity to improve productivity in terms of outputs. Second, as regards its contribution to the effectiveness of public policies – or in other words, its ability to produce the outcomes (final objectives) targeted by these policies and which are often linked to social wellbeing or economic growth.

After examining the place that indicators now hold in public sector decision-making and identifying their properties, virtues and perverse effects, the article goes on to analyse their performance in terms of the efficiency of public organisations and the effectiveness of public policy. It concludes by listing the conditions for an appropriate use of indicators in public sector management and decision-making.

4.1. A technology of government based on indicators[16]

As the former U.S. Secretary of Defense, Robert McNamara commented "You cannot make decisions simply by asking yourself whether something might be nice to have. You have to make a judgment on how much is enough"[17] (quoted in Enthoven and Smith, 2005). In 1963, McNamara applied the investment selection process he had designed for managing the Ford Corporation (Wildavsky, 1969; Thoenig, 1971) to public sector management and set up a so-called rational approach to policymaking, which he dubbed PPBS (Planning, Programming and Budgeting System)[18] (Mallard *et al.*, 2009). The Cold War context offered fertile ground for the passions and bureaucratic power games likely to influence military affairs. His aim was thus to escape the clutches of the various lobbies by rationally controlling – i.e. through an impersonal and methodical approach – the link between the budgetary resources allocated to his department and to its weapon-building programmes. Decisions were to be made using explicit criteria of national interest; needs and costs would be considered simultaneously, decisions would be taken after a candid confrontation of possible alternatives; they would be based on relevant and non-partisan data permitting *ex ante* programme evaluation, mainly using economic calculations. When the first budgetary crises loomed in the 1960s with the ambitious «Great Society» programme and the expansion of the Vietnam War, President Johnson extended the procedure to the entire federal administration. What later proved to be a technocratic utopia migrated towards Europe in the late 1960s. Such, for instance, was the origin of the Rationalisation of Budget Choices (RCB) approach promoted by Michel Debré and the French Ministry of Finance.[19]

At the same time, various social scientists were proposing to develop indicator-based analyses in order to compare the pay-offs of public policy alternatives and to "give a voice" to citizens' needs. This trend took on strong momentum in the United States with the development of experimental policies (housing, redistribution, negative tax, bussing, etc.). In the 1980s, propelled by the

[16] I thank Jean-Claude Thoenig and Vincent Spenlehauer, who will each recognise their generous and invaluable contributions to the writing of this paper.

[17] Robert S. McNamara, Remarks to the American Society of Newspaper Editors, Washington, DC., April 20, 1963 and as DoD Press Release, No. 548-63, quoted in Enthoven and Smith, 2005.

[18] The US Department of Defense (DoD) Planning, Programming and Budgeting System (PPBS), 2004. A Historical Perspective, 37th Annual DoD Cost Analysis Symposium.

[19] For recent developments of this tool, see for example Perret, 2006.

same trend, France extended what government planning had set up with the help of a National Accounting system (Fourquet, 1980) to the analysis of social needs in a welfare-type society. This movement was driven by influential experts and public servants such as the well-known economist and politician Jacques Delors.

This brief historical review shows what drives these innovations in the management of public affairs: *indicators play a key role as tools for steering organisations and public policy, and become legitimate public choice criteria as functional elements of a government system. They supply the basis of a new technology for government*.

Gösta Esping-Andersen (1990) has shown that until the 1980s, post-war European welfare states had injected massive amounts of resources into their social protection systems without undertaking any cost-impact analyses, as if merely implementing such policies would automatically turn their intended goals into actual outcomes. Facts repeatedly proved that this was not the case, that the impacts generated were not in line with the goals defined by the policy-makers. As pressure for public intervention grew faster than the resources available, the rise of neo-liberal ideas began to suggest that the ineffectiveness of public policies stemmed from the inefficiencies of government bureaucracies. There was insistence on the need for an alternative management method that could reduce public sector costs by calling on quasi-market models of regulation and delivery. This approach would also strengthen the links between policy orientations and organisational activities. By decentralising and bringing micro-management as close as possible to operational units and by transforming bureaucratic injunctions into quasi-market incentives, the policy-steering system would delegate the evaluation of policy implementation to end-users that enjoyed drawing rights enabling them to express their preferences. This would also upgrade the quality of services provided by such institutions as the cumbersome and costly hierarchical control by public authorities would be eliminated and competition enhanced.

The properties of a quasi-market management model are in sharp contrast with traditional public administration mechanisms, which in fact combine political, compromise-based decisions and bureaucratic rule-based implementation.

(a) It replaces hierarchical subordination with accountable autonomy. Autonomy means the capacity to make strategic decisions in an environment that offers specific resources and constraints. Accountability refers to the fact that one is able to account for one's actions in terms of their appropriateness and results compared against the orientations defined by the policy.[20]

(b) The model requires a deep-cutting restructuring of the administrative organisation, in which the hierarchical subordination of devolved services is

replaced by the broad autonomy of operational organisations. It means globalising and transferring to the latter all the resources (human and budget resources, movable and immovable assets) previously managed on a silo-by-silo basis by the central administration; embedding their strategic capacity in cost accounting and management control tools; and putting an end to the hegemony of the *a priori* control assigned to the supervisory authorities.

(c) The model entails an entirely different conception of management and control by public authorities. If the self-government of public organisations is to be promoted, internal governance needs overhauling so that strategic competence and operational authority are based on local leadership rather than on hierarchical or professional status alone.

(d) The model completely redesigns the relationship between public operational organisations and public authorities. It replaces *ex ante* compliance with formal regulation by *ex post* assessment of the actions taken with regard to prescribed policy objectives. Thus, the organisation's "performance" is understood as the controllable result of its action (a product, an output), rewarded or penalised through the resources allocated *ex post* by its stakeholders, among which are its supervisory authorities. A cybernetic loop links incentive, performance evaluation and reward/sanction. In this model, public authorities become a *principal* able to remotely govern *agents* that contribute to policy implementation. It deals with public policy steering as a problem of resource allocation meeting two objectives: to enhance the effectiveness of policies by increasing their influence on the practices of public organisations and to render the expenditure of public organisations more efficient by streamlining administrative practices.

(e) Last but not least, this approach requires operational tools to assess the quantity and quality of what are usually non-market products delivered by public organisations through public policies. This is the role of performance indicators. They link policymaking to implementation through quasi-prices, *i.e.* measurable and credible proxies for the contribution of organisations with respect to the set policy objectives. Evidence-based management – like evidence-based policy – provides the basis for controlling and assessing organisations and policies using the objectivity of quasi-prices that are supposed to benchmark and sanction performance.

This type of framework associates a technical substrate and a management philosophy based on a simplified vision of organisational relationships between policy designers, users, advisers and overseers (Hatchuel and Weil, 1995). *Combining the values of transparency and accountability, this technology replaces direct injunctions with mechanisms that make the behaviours of organisational actors computable by systematically influencing the conditions of their actions,* thus rooting a computational rationality into

their way of operating (Altfeld and Miller, 1984; Miller, 2001).

Hence, two questions are raised: one regarding the efficiency or the capacity of "indicator-based management" to improve public organisations' productivity in terms of outputs; the other regarding their contribution to the effectiveness of public policy or their capacity to produce the outcomes (final objectives) defined by a specific policy, which are often related to social wellbeing or economic growth.

4.2. The efficiency of public organisations: the role of indicators.

Where does the belief in the value of indicators for generating better public management and ensuring more efficient policy enforcement and implementation come from?

4.2.1. Rigour and impersonality

Figures play an increasing role in public decision-making. They define and assess the objects they identify in a more uniform, accurate and rigorous manner. This explains their success and proliferation in the form of various indicators that mobilise a multitude of evaluation tools (certification, standardisation, management control, rankings...).

The objectivity of indicators derives from the rigour of numbers, in the sense that it sets the objects to be assessed at a distance from the pressures, prejudices, passions and vested interests of their stakeholders (Porter 1995). Indicators provide uniformly applied measures according to references that in principle are known or knowable by all. The credibility of indicators depends on the extent to which they are correlated with external rules that secure the independence of the agencies operating them, and with internal rules that ensure the organisational and scientific quality of the production systems providing them. Alain Desrosières (1993) highlights the huge and time-consuming work needed to build the institutions, knowledge, techniques and professional bodies that create a collective trust in the impersonality and honesty of statistical measurement. *The trustworthiness of indicators depends upon how much their users trust the accuracy of the techniques, bodies and institutions that produce them.*

There is little risk that established indicators will be manipulated, as their computation is "disciplined" by the routine operations that produce them. Moreover, the fact that they are closely interlinked with other indicators also requires them to be durably coherent, which discourages the chains of complicity operative in cheating practices.[21] "Those who think they can manipulate numbers at will are often proved wrong" (Espeland and Stevens, 1998: 331). Large-scale indicator-building processes, such as the Bologna Process in higher education, demonstrate that national actors who are regularly asked to inform a battery of indicators may be initially tempted to present over-flattering images of their situation. However, they are soon caught up by the unremitting requirement for the internal coherence of their scoreboards (Ravinet, 2011).

[21] Even though this is not excluded, as is often shown by the "manipulation" of unemployment figures, which generally entails changing the definition of an "unemployed worker", or by the more exceptional but tragic example of fraud in the production of the Greek government's accounts.

Indicators thus have mechanical merits: they provide a technically unbiased proxy for the objects that they bring into existence. In this respect, they play the role of referee and provide a basis for comparison: anyone can contest the validity of the indicator used to describe a particular object, but no one can challenge the result produced by the measurement process, provided the process itself is trustworthy. Indicators thus have the capacity to reveal realities that are hidden by social representations. In this respect, they constitute powerful tools often much appreciated by challengers, whose achievements may go unnoticed when they are overshadowed by the prevalence of established and sometimes overvalued reputations. A typical example of this situation is the increasing use of performance indicators in higher education, since these can direct the spotlight onto ambitious and successful universities that nonetheless lack the "noblesse" needed to show up on the radars of social prestige. Indicators draw their strength from the fact that they conventionally extirpate singular entities from their incommensurability and group them into categories of similar objects that make comparison possible (Karpik 2010). "Commensuration offers an adaptive, broadly legitimate device for conferring a formal parity [to objects] in an unequal world" (Espeland and Stevens, 1998: 330) (see also Espeland and Sauder, 2007). *Indicators de-contextualise the objects they measure by creating uniform terms of measurement that disregard the time, space or culture in which they are embedded.*

Classification faces a recurrent dilemma. For the sake of comparison, numbers reduce complexity by disregarding what the entities they capture owe to their temporal and spatial context. In doing this, they may discard information that could be useful for gaining a better understanding of their behaviour. The art of building and using quantitative indicators is the art of playing with data that *translate and, in this respect, betray the complexity of so-called reality by creating, approximating or distorting it. Both the producers of indicators and their users come up against this issue.* Bibliometric indexes provide an obvious illustration. They supposedly pave the way for comparing scientific performance across a set of authors, disciplines or institutions in that they count the number of publications or citations over a given period. Nevertheless, the methods they use to qualify and standardise performance measurement fail to take account of the fact that quality standards across disciplines vary enormously in terms of time to publication, length of papers, diversity of publishing supports or the rate of citation accrual, durability of citations, and so on. Bibliometric indexes thus betray singularity by imposing exogenous assessment criteria on performance in the various disciplines.

One last point worth underscoring is the potential for comparison incorporated into an indicator. This depends on its durability and scope within changing social environments. Indicators aim to make sense of the realities they account for, while at the same time comparing situations that are distant in time and space. As a result, indicator designers are anxious to build plausible compromises between adapting their indicators to current realities and maintaining the continuity of a time series or the scope of observation, as the example of

Table 2 *Publication practices by scientific discipline*[22]

Publication practices	Maths	Chemistry	Physics	Biology	Sociology
Preferred publication vectors	Memoires	Journals	Journals	Journals	Books
Perceived rank of the journal	Strong	Strong	Strong	Variety	Variety
Community consensus on the perceived rank	Strong	Strong	Strong	Strong	Rather weak
Time to publication	Slow	Average	Average	Average	Average
Accrual of citations	Very slow	Quite fast	Quite fast	Quite fast	Quite fast
Lifespan of citations	Long	Quite short	Quite short	Quite short	Long
Collaboration	Quite rare	Very common	Very common	Very common	Few, increasing
Co-signed PhD supervisor / students	No	Yes, not a rule	Yes, not a rule	Yes, a rule	No, increasing
Publication rhythm	Rare	Frequent	Frequent	Frequent	Quite frequent
Length of papers	Long	Short	Short	Short	Average

the French INSEE's occupational categories well illustrates (Desrosières and Thévenot, 1988).

4.2.2. Proxies and content: adverse effects

Indicators are naturalised proxies for a reality that they, in fact, help to create. They qualify the properties of an institution or a person in a simplified manner, usually along an ordinal scale. As such, they are likely to induce unexpected effects that are now well-known."[M]easures are *reactive*. Measures elicit responses from people who intervene in the objects they measure" (Espeland and Sauder, 2007: 2).

Indicators alter expectations and consequently behaviours with respect to the objects they measure. Published rankings of universities, hospitals, companies, countries, etc., based on surveys either by public authorities, independent agencies or private actors such as the media, substantially

[22] Reviewed by C. Paradeise, based on Schlenker, 2009.

influence how these are perceived by external audiences (Paradeise, 2012). Their public tends to opt for the highest-rated institutions without checking how far these fit their specific needs. Since available resources are increasingly captured by the better ranked, institutions tend to ground their strategies on those indicators likely to enhance their ranking. By doing so, they thus reinforce the validity of the measurement and generate self-fulfilling prophesies.

Being exposed to indicators *alters our cognition* (Espeland and Sauder, 2007). By transforming qualities into quantities that share the same metrics, they impact the way in which we pay attention to the world around us, the connections we establish between entities and the manner in which we express similarities and differences: they naturalise social forms and make it difficult to express and convey realities that are not covered by their metrics.

By impacting how the value of an entity is perceived and assessed, *the use of indicators redistributes resources, redefines work and modifies professional values and standards.* A bank manager, just like a local policeman, is required by his/ her hierarchy to "reach his/her quotas". This makes local interpretation and appropriation of occupational rules and duties difficult if not impossible, and thus impacts the structures of authority, sociality and responsibility, the meaning of work, and the very notion of what is to be valued. The case of universities (as well as hospitals, schools, police services, courts, etc.) being suddenly exposed to the iron rule of rankings provides a simple illustration of such phenomena.

A number of papers have described how the access to public and private resources (budgets, contracts, students and qualified teachers, etc.), organisational governance, division of labour, employee evaluations, etc. are simultaneously impacted by the development of rankings that combine baskets of indicators. Indeed, valuing organisations or individuals according to their ranking on a basket of indicators leads to a remote style of government that encourages "organisational isomorphism" (Di Maggio and Powell, 1983) – *i.e.* encourages conformity and the alignment of organisational perspectives and practices with the model prioritised by the indicators used. Remote principals steer the behaviours of distant agents by using incentives, the impact of which in terms of performance is measured by indicators. Indicators govern how principals reward and sanction agents and hence "discipline" agents' behaviour. But this approach can hold serious risks. It may prompt organisations to focus their strategy on very short-term performance indicators and reduce their portfolio of missions by over-focussing on the highest pay-outs. This may have the unfortunate consequence of driving universities to concentrate on research rather than teaching, police stations to focus on repressing petty crime rather than community work, bank employees to promote the latest financial products rather than paying attention to the profiles and needs of their clients, etc.

The pressure to conform becomes a threat for organisations, as these run the risk of paying more attention to form and forgetting about substance and content (Merton, 1940). Such pressure often creates *decision-making processes whose purpose is*

to "play with the rules". Reducing activity to performance-based indicators may lead to confusion between "content" and "content signals". In order to "fit» with the leading indicators for major international rankings, a university may prefer to pick a Nobel prize winner from a star-spangled labour pool so as to inflate its bibliometric and reputation indicators, rather than patiently «cultivate» its own academic labour force; to please shareholders, a company may favour rationalisation and lay-offs, thus signalling its concern for short-term cost efficiency; or a police department may "rack up arrests" and discourage citizens from formalising complaints, etc.

Indicator-induced tyranny may also impact initiative and innovation, and crush entrepreneurial spirit and professional commitment. *The disadvantage of systematically using the mechanical objectivity of indicators is that it dismisses the contribution that actors make to an organisation in the form of experience. Indicators create "rituals of compliance" that discourage patient attentiveness to the sources of deficiency and success* (Power, 1997). Indicators exert a *soft constraint* on organisations (Courpasson, 2000); no one knows why and how top-down instructions are given, or by whom, which thus weakens the autonomy of middle management; options are defined by automated routines that discourage people from paying attention to "weak signals"; decision-makers cannot be identified or responsibilities attributed. Altogether, these drifts lead to dangerous losses of professional commitment and drain leadership potential.

In addition, subservience to indicators creates the risk of an *"arms race"* within winner-takes-all markets (Frank and Cook, 1995), when – as in a sports tournament – the winner comes away with the jackpot while the second or third best players are awarded little more than consolation prizes, even though the gap between the gold and the other medals may be infinitesimal. *In other words, the star system applied to life in society means that what is rewarded is not simply being "good" at one's activity but being "the best".* Hence, the temptation to "fit with the indicators" that calibrate the gains distributed. This "arms race" comes about because, in this type of championship, everyone is trying to improve their overall position by acting on all the indicators that help define it, including the indicators the furthest away from the core activity. Some authors have drawn attention to the harmful effects that these policies have on American universities (Ehrenberg, 2000; Sperber, 2000; Clotfelter, 2011): enrolment fees are increased to fund the most lavish campus and to recruit and subsidise the best soccer team, since the choices of students and their families (as well as patrons, alumni and public authorities) are primarily based on rankings. This pushes universities to incur exorbitant expenses for lawns, gyms, etc., whereas the collective benefit from such facilities is quite marginal in the light of their missions of education and research. *When everyone is racing to reach the same goals or to fit the profile, the result is a massive increase in the collective cost of this effort, whereas the collective benefits are limited with respect to the missions assigned to the organisations bearing this cost.*

4.2.3. The dual nature of indicators in the management of organisations

Monitoring efficiency is an internal function of an organisation. It assesses the gap between a desired (or average) performance and actual performance conventionally evaluated by indicators. Indicators thus play a dual role in the organisation. Internally, they can be used to allocate resources according to a mechanical principle, such that the consequences of poor performance evaluations are not subject to pressures from inside the organisation. They are also more widely used as tools to diagnose the organisation's strengths and weaknesses, and analyse the data and processes that help to explain these. As such, indicators provide support for strategic organisational decision-making. Externally, indicators provide the supervisory authorities (central administration for a public entity and head office for a private entity) with an *ex-post* evaluation of performance that helps them determine their overall policy and specific strategies for their cost and profit centres.

The decision-making and implementing processes of an organisation, whether private or public, are never accurately reflected in the picture suggested by its official organisation chart and line of command. In an organised environment (polity), steering human action – which includes political power dynamics – is not like operating a machine. Action takes place within space-time continuums, and this requires being able to make plausible diagnostics and reasonable anticipations when faced with alternatives, at the same time taking all the actors involved into account. When top management levels announce a given action or reform, their internal constituencies do not only (or even first) consider the rationality of the proposal for the company as a whole. They (also) consider the consequences that such decisions may have for themselves and the impact it could have on the way they work. Many well-intentioned reforms may not succeed because senior echelons have underestimated the resistance to change at the middle and lower levels of the organisation. The latter may be reluctant to implement changes that they think will disrupt the existing internal order and threaten units with strong power positions! At best, an organisation usually moves two steps forward and one step back. Detailed reform designs and authoritarian decision-making processes often fail. A McKinsey study (Fubini *et al.*, 2006) shows for instance that half to two-thirds of company mergers fail after a year[23] because one of the buyer's first moves is to impose a rigid plan that under-utilises opportunities and hurts and humiliates the acquired company, etc. *Polities are polyarchies*, which is to say that they cannot be reduced to worlds organised by a single principle of authority, but are based on the interplay between participation and protest (Dahl, 1961). Efficiency in terms of internal organisational change, as Lindblom (1959) describes it, is much more likely to be fuelled by adopting a "disjointed incrementalist" approach – a series of small-

[23] According to a generally accepted criterion, a merger is considered successful if the stock exchange price of the new entity performs better than the average in its sector.

scale unconnected actions – than by adhering to a carefully planned approach that focuses on the sequence of its operational details. This holds even more true for the professional bureaucracies (Mintzberg, 1979) typical of public sector organisations. In these "loosely coupled" (Weick, 1976) and "heterarchical"[24] worlds (Stark, 2009), professionals enjoy the autonomy deemed necessary for the satisfactory organisational performance of their missions. They resist the injunctions of a managerial hierarchy or quasi-market incentives, which they usually reject as being inappropriate or illegitimate with respect to their missions.

Ultimately, what public sector organisations discover as their internal indicator-based management develops is what private organisations discovered a long time ago when introducing management control, *namely, that indicators alone are not sufficient. They produce both virtuous and adverse effects depending on whether they are used mechanically or with reflective understanding.* Later in this paper, we shall list several prerequisites to prevent dysfunctional consequences and foster virtuous effects.

[24] By «heterarchy», Stark (2009) refers to polities characterised by the plurality of values that are contributed by their different components.

4.3. Indicators and policy effectiveness

Does the quest for organisational efficiency suffice to secure more effective public policies? Experience repeatedly suggests that this is not the case.

Our societies too often consider that the right information – information that delivers an unbiased and unambiguous understanding of reality – can be synthesised in the form of a ratio or score. Moreover, the onset of remote government has accentuated this perception by reinforcing the idea that the effectiveness of a public policy is equal to the efficiency of its implementation by the organisations in charge. Without falling into a purely technocratic vision of what a policy means, there are good grounds to question this double confusion: first, the objectivity of figures does not mean that they are consensually viewed as the ultimate reference or that they are not open to debate; and second, the effectiveness of a policy is not equal to the efficiency of the public organisations officially accountable for its implementation.

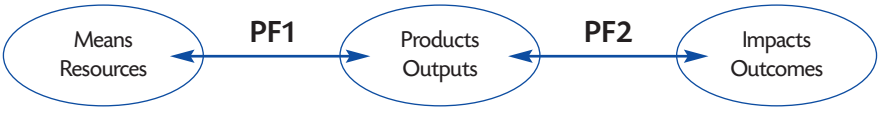
4.3.1. Efficiency and effectiveness

Seasoned evaluators of public policies rightly make a basic distinction between two concepts. On the one hand, they analyse outputs, which relate to the performance or the *efficiency* of the organisations in charge of implementing them. On the other hand, they consider outcomes that serve as proxies for the *effectiveness* of a specific policy, as for instance its societal impacts. O*utputs* of an education system can, for example, be measured by the percentage of students of a given age that obtain a given qualification, while its *outcomes* measure the impacts that the resources mobilised by this policy induce on the workforce's level of qualification and its contribution to social wellbeing (Mandl *et al.*, 2008). The two concepts of outputs and outcomes refer to two clearly distinct "production functions" (Gibert and Andrault, 1984; Meny and Thoenig, 1989).

In the diagram below, PF1 refers to indicators that link the mobilisation of means and resources to the delivery of outputs such as products and services. Efficiency is therefore assumed to be internally generated and administered by the organisation in charge of implementing a policy, in conventional terms such as the quantities delivered, their costs and their quality. The organisation is acting as an agent that is required to comply with a mandate given by a principal, the policy-maker. It is not responsible for defining the reasons and purposes of a policy, simply for enforcing it. Efficiency is monitored by verifying the effectiveness and reliability of the organisations responsible for implementing a policy that has been defined by legitimate public authorities (public or private organisations, under state control or subcontracted). It disregards the effectiveness and reliability of the actual

policy and is essentially focused on the implementation phase. The monitoring of efficiency generally relies on indicators that measure the volumes of direct outputs achieved by the implementing organisations compared to the resources consumed.

PF2 refers to the way such outputs produce (or not) specific or desired societal outcomes and impacts on the societal fabric. At this level, however, the policy's effectiveness may be perceived differently by the various stakeholders concerned to the extent that they may not share the same judgments. Effectiveness indicators may thus vary because the values, interests and perceptions are not conventionally identical.



It quite often happens in real life that these two production functions are confused. Efficiency indicators are considered as effectiveness indicators or vice-versa. When a policy is failing in terms of outcomes, to make it more successful, attention and initiatives may then focus primarily on upgrading the internal efficiency of the operators implementing it, while the fact the policy may have failed for other reasons is overlooked; failure may be due to a policy design aimed at producing societal changes that are impossible to bring about, or because the outputs defined do not generate the intended outcomes, or the expected outcomes do not occur because society and outside stakeholders see them as non-desirable or threatening.

A policy evaluation seeks to assess which results or external impacts can be attributed to its action by moving up the chain of hypothetical relationships that link it to the policy outcomes, as in a chain of cause and effect, and which go far beyond the measure of how efficiently agents are implementing the policy. This explains why it can be argued that the outcome cannot be observed by cost or quality indexes. The estimation of outcomes derives from an intellectual reconstitution comparing an observable situation at a given time following the implementation of a policy to the hypothetical situation that would have been observable if this policy had not been decided and implemented (Gibert 2003).

Evaluating policy effectiveness thus has some commonalities with monitoring the efficiency of the organisations that help to implement it: in both cases, this involves identifying gaps in order to interpret them in light of a reference value and adopting tools to try and eliminate biases using concrete evidence. But what distinguishes the evaluation of policy effectiveness is its *ad hoc* nature, which prevents its becoming a routine operation, whereas management control or efficiency monitoring derives its strength from its systematicity and regularity.

4.3.2. Measuring efficiency and effectiveness: an arena open to pluralistic debate

Policy evaluation is always designed and run on an *ad hoc* basis as it addresses situations

that are non-reproducible. It seeks to identify the implicit theory of social order and dynamics on which the policy is built, not only to explain its potential benefits, the form it takes and its fit with the social order it intends to impact, but also to delimit the conditions conducive to its success.

This theory of action can be apprehended through the indicators selected by the policy in order to express the outputs of the implementing organisations. To assume that an indicator is objective or universal in no way means that it expresses the reality of things or a natural truth (Porter, 1998). Different economic, aesthetic and moral values coexist in society at any given time (Espeland and Sauder, 2007) and everyone uses their own yardstick to estimate the value of an object or an event (Boltanski and Thevenot, 1991; Stark, 2009). As a result, the same object may be valued in diverging and incommensurable ways. An indicator thus expresses a biased view of reality, as it translates a representation of reality into a scale of values (by evaluating human life in terms of price, the quality of a university through the "excellence" of its academic publications, an ecological disaster in terms of monetary damages). It thus imposes a specific vision of the intrinsic characteristics of the objects that it helps to build. Moreover, the priority accorded to the valuation principle in evaluations crowds out other possible principles. One example is GDP, which is the common yardstick for describing the wealth produced by a nation. The fact that GDP takes no account of measurements of wellbeing or the externalities produced by growth was challenged in the 1970s by the Club of Rome and by the "social indicators" movement (Delors, 1971) among others; the Stiglitz Commission[25] was also set up in France in 2008 to seek alternative methods for wealth valuation.

However naturalised they may be, indicators elaborate and embody conventional and socially constructed visions of what reality means. For example, Porter (1995) emphasises the fact that "society" is a construction induced by all the measures that characterise a territorial space taken as a reference. Bourguet (1988) shows how the "reality" of the French nation was invented at the turn of the nineteenth century through the administration's trial-and-error efforts to rationalise a group of disparate territories by the trial-and-error application of indicators that may today sometimes seem somewhat absurd. Research by Desrosières and Thévenot (2006 [1988]) describes the singular statistical history of how occupational categories were established in France. They emphasise how the evolving social structure obliges statisticians to make compromises between the need to stabilise these categories for the purposes of inter-temporal comparison and the need to redesign them so as to capture the reality they are trying to describe. Musselin (2001) shows that the fact that French universities have been obliged since the 1980s to draw up a four-year indicator-backed project has continued to play an essential role in shaping a territorial identity they had ceased to enjoy since their dismantling during the Napoleonic period.

[25] Cf. http://www.stiglitz-sen-fitoussi.fr/en/index.htm

In other words, indicator-based quantification is "politicised": Rose (1991) stresses that political judgments are implicit in the choice of what to measure, how to measure it, how often to measure it, and how to present and interpret the results (1991). As naturalised cultural objects (Desrosières, 1993), indicators have the advantage of not being challengeable by their stakeholders. They are assumed to offer a valid representation of the reality that they claim to explain, thereby justifying the fact that no attention is paid to what lies outside the scope of what they measure. Their institutionalisation creates a barrier to seriously envisaging that alternative realities may exist and matter and, therefore, that other ways of partitioning reality and describing it with numbers may be just as relevant, or that other facets of a situation could be usefully taken into account. As they are crystallised within a set of technical mechanisms, indicators appear "self-evident", because our visions of political reality are shaped by what statistics appear to disclose.

When indicators become a stable, naturalised part of their environment, experts and policy-makers forget that they were in fact *built*; that the conditions in which they were designed and developed matter (when, in what contexts they were originally set up, and for which specific purposes); that before gaining acceptance and the status of objective technical tools, the debates surrounding them were very often heated; and that they gave rise to conflict and competition between opponents and alternative views. As crystallised products of the power of expression – and of the competition this may create –, indicators tend to carry visions of reality that are generally confused with "reality" itself. They are only revealed in their true light when put to the test by concerned groups who publicly and vociferously challenge the values they embody. Indicators therefore could be considered as a kind of glossary of a social order. When they are challenged, it is a sign that this order is exposed to destabilisation.[26]

4.3.3. Evaluation of the effectiveness of public policies and indicator-based measurement

Public policies fulfil a function that is external to the organisations implementing them: their aim is to change a current state of affairs or to avert any threat to situations judged as satisfactory. A policy is a theory of change insofar as it assumes that a relationship exists between the outputs it delivers and the outcomes or impacts these induce. The problem lies in the fact that this causal relationship does not exist in a vacuum and that cause-effect relationships are non-linear. A policy operates in complex societal contexts where it has to integrate

[26] Granet (1988 [1934]: 363) refers to the fact that when Chinese feudalism found itself weakened, schools of thought flourished with the aim getting things back into order by correcting language. He cites Confucius' response to the question: "'What will you consider the first thing to be done?', the Master [Confucius] replied, "What is necessary is to rectify names... If names be not correct, language is not in accordance with the truth of things, affairs cannot be carried on to success... punishments will not be properly awarded... the people do not know how to move hand or foot; Therefore, a superior man considers it necessary that the names he uses may be spoken appropriately, and also that what he speaks may be carried out appropriately. What the superior man requires is just that in his words there may be nothing incorrect" (Confucius, The Analects, Translation by D.C. Lau, cited in Granet,1988 [1934]: 363).

multiple sources of complexity. It cannot make a clean sweep of existing policies incorporated into established mechanisms. It triggers reactions from a diversity of actors – within and outside the public sector organisations in charge of its implementation – who view it as either a resource or a constraint with respect to furthering their own interests. It cannot disregard the many influences that come to bear on a given policy. Neither can it ignore the contradictions that exist between the objectives of different policies pursuing the same targets – for example, the restrictions on foreign students attending French universities, introduced by the French government early 2012, were caught in the crossfire between immigration policies, skills shortages in some professions and international relations.

As key items in the policy makers' tool kit, indicators have also become key items in the approaches to public policy evaluation. Statistical and econometric analysis is often used to characterise the impacts of a specific policy, either *ex ante* in the form of experimentation prior to scaling-up, or *ex post*. However, it is this type of analysis that finds it difficult to elucidate the processes that give rise to policy impacts. Such highly macroscopic approaches can only express what they incorporated into their measurement mechanism at the outset, either by choice or by necessity given that relevant existing data is usually lacking. They are therefore often criticised for looking for the key under the street lamp and too hastily locking the evaluation into a hypothetico-deductive methodology. Their advocates disdain the inductive approaches used by the social sciences, which on the contrary open up avenues to an in-depth interpretation of the gaps between expected and observed outcomes. Indeed, inductive approaches make it possible to test the coherence of credible scenarios in which observed results can be attributed to the policy evaluated, while also taking account of the behaviour of the actors that are affected in different ways by the policy outputs within complex and evolving systems.

Indeed, every policy is subject to debate regarding its formulation, the sources of dissatisfaction and problems, as well as the opportunity of putting it on the public agenda (how efficiently agents are implementing the policy) (Gibert, 2003). *This should be kept in mind. Otherwise, there is a strong likelihood that the evaluation will be mistakenly seen as a form of internal control, which it is not. Evaluation is a judgment on the appropriateness of a decision with regard to principles that are deemed legitimate by the policy's multiple stakeholders. As a result, impact indicators constitute only one of the many building blocks that help to build the judgment.*

Evaluation is thus not compatible with a ballistic vision of the relationships between policy as a cause and impact as a consequence, as portrayed through macro-econometric lenses. Implementing an effective policy is in no way analogous to sending a rocket to the moon at an acceptable cost. In the latter case, all the relevant scientific laws (astronomical, technological, physical, etc.) are known, no independent external actors intervene in the process and the organisation leading the project is in full control of events, even though it may also make tragic organisational

mistakes. Yet, policy makers very often address a complex social situation as if it were a ballistic planning process. One of many examples is the *a priori* experimental evaluation of a new policy that was adopted in France to provide unemployment compensation based on an "active solidarity income" (RSA). This demonstrated in an (almost) technically perfect manner[27] ("with a result statistically significant at the 5% level") that the new mechanism would lead to a 30% improvement over the poor performance of the former scheme providing "minimum income benefits" (RMI), even increasing it from 2.25% in the experiment control areas to 2.92% in the test areas (Barbier, 2011). As Spenlehauer (2011) points out, the actual relative ineffectiveness of this policy could have been anticipated "by looking at actors with their complex structure of strategic objectives and behaviour, their possible theories for action and their changing inter-linkages"[28] (Spenlehauer, 2011: 199). He cites the example of small employers in France, for whom the shift from the RMI to a conditional RSA was far from neutral. This was also the case for the local public employment services (Pôle Emploi), which were being set up at the time following the decision to merge the two former national employment agencies (the ANPE and ASSEDIC).

Societies, organisations and polities are stubborn enough to resist the wishes of technocrats! Evaluation must be recognised as an exercise that cannot be reduced to mere technology or an abstract methodology that deduces impacts by calculating indicators. The first reason for this is that evaluation cannot argue that the objects it calls "policies" always (or even generally) have clearly defined objectives in the sense of specific and unambiguous goals that can be readily analysed *ex post* to show whether these had succeeded or failed. In fact, "policies" tend to cover a sequence of actions whose goals are not necessarily very clear, fluctuate and prove contradictory over time. The second reason is that, although evaluators endeavour to obtain evidence-based results, they cannot ignore the fact that public policies are embedded in stratified, discontinuous, conflicting and siloed systems. The systems themselves are riddled with ambiguous relationships in which no consensus exists on action agendas, on the way issues are defined or on policy design. They comprise a multitude of open windows offering both resources and constraints.

As a result, evaluation cannot be limited to simply monitoring the efficiency of public organisations with regard to the objectives assigned to them by public authorities. The evaluative judgment targets the impacts of a policy, and thus the quality, value, appropriateness and theory of change underlying the policy with respect to its fixed objectives. How, for instance, can the death penalty be evaluated as a criminal policy and on what premises – factual or normative – is

[27] Bourguignon F, *Note de synthèse*, in *Rapport de synthèse sur l'évaluation des expérimentations RSA, Comité d'évaluation des expérimentations*, Septembre 2008. The complete *Assemblée Nationale* report (session of 25.9.2008) can be consulted at: http://www.assemblee-nationale.fr/13/cri/2007-2008-extra2/20082010.asp.

[28] Author's translation

it based? The answer differs across polities, societies and epochs. Evaluating this policy means assessing the impacts of its built-in assumptions by analysing how its implementation impacts the normative and factual objectives it is targeting – not "to attain a quota" of people sentenced to death, but to reduce public insecurity, for example. Thus, effective evaluation would necessitate a rigorous analysis of the induced effects of the actions implemented under the banner of the policy being assessed. Evaluators need to understand how the policy's resources or constraints are appropriated by society at large by examining how the *targeted actors are impacted by the tools it provides them with*, and thus by rejecting the assumption that policy implementation is a linear and mechanical consequence of policy design and of the stated intentions of the policy makers.

Viewed from this angle, indicators unveil problems or surprises – the gaps between expected impacts and the actual impacts measured – but they do not clarify precisely why and where these gaps occur. In other words, they may help calculate statistical correlations but they offer no ready-made explanations. Other approaches are needed to interpret and explain the causal linkages. Road safety policy provides a good illustration of this need. The statistical data on the number of road deaths in France show that after a moderate decline in fatal accidents over several years, there was a sudden spectacular decrease following the installation of automatic radars in 2002-2003.

But how are these results to be interpreted, given that other measures with less spectacular impacts had regularly been put in place over the course of many years[29] to deter reckless driving? Does it have something to do with the aversion to discipline that supposedly typifies "Latin folk"? Or is there a connection with the perceived social status that high-speed cars hold for certain social groups? Or is it linked to the fact that paying a fine costs less than the time lost by respecting speed limits or the regulatory driving hours for truck drivers (Ocqueteau and Thoenig, 1997)? Or to the fact that a driver estimates the degree of danger differently depending on the driving situation (proximity, long-distance)? Policies based on deterrence (speed limits), on sanctions (points-based driving licences, breathalysers), or on communication and education (changing attitudes to short local journeys, education, awareness-raising for children) do prove moderately effective. Yet, what proportion of these results can be attributed to road safety policies as opposed to road improvement works, the manufacturers' enhanced vehicle safety or a higher level of education, etc.? The answer to what exactly caused this sudden increase following the installation of speed radars is far from obvious.

A policy evaluation carried out in 2002 reformulated both the question and the problem. It suggested that previous policies had disregarded one key factor, namely the behaviour of the public authorities (police, courts) in charge of implementing the sanctions. International comparisons show

[29] 1973-90: speed limitation and extension of areas covered; 1978-95: law to prevent drunk driving with increasingly lower thresholds; 1979-1990: widespread use of safety belts; 1992: points system on driving licenses; 2002: road safety a major national campaign and beginning of the installation of radars.

that in countries where specialised police units are responsible for implementing road safety, the results in terms of impacts and driving behaviours are more rapid, spectacular and stable than in countries where they are handled by the general police force (Ferret and Spenlehauer, 2008). Why is this? Because a general police force has to deal with matters relating to between 80 and 100 different policies (transport, immigration, petty thieving, criminal investigations, etc.). The police rank their actions in terms of priorities according to a hierarchy of prestige that relegates road-safety checks to the bottom of the list. Community police officers are also quite sensitive to the potentially serious consequences of inflicting huge sanctions on self-employed and economically fragile carriers and truckers who are geographically and socially close to them. In light of this analysis, and given that no specialised road police units exist in France, or are even be likely to be created, the Ternier Report (Phélippeau, 2003) proposed that roads be massively equipped with automatic radar systems to detect and sanction speeding. This has not only considerably reduced the leeway for individual and local arrangements for policy enforcement... but also substantially impacted user behaviours.

4.4. Some recommendations for practice

So, how can indicators be usefully exploited? Looking back at his experience as former dissident and president of his country, Vaclav Havel made the following comments:

> I was succumbing to the kind of very destructive impatience that is characteristic to modern technocratic society, imbued with its rationality, persuaded wrongly that the world is nothing more than a crossword puzzle that had only one correct solution – a so-called objective solution – to the problem. Without my realizing it, I was succumbing de facto to the perverse certainty that I was the absolute master of reality – the master whose only vocation would be to achieve this reality in accordance with a ready-made formula ... The fact is that the world, Life and history are governed by a time that is uniquely theirs ... ... To want to suppress this impenetrable "tortuosity" with an infernal barrage entails plenty of risk ... I had wanted to move history forward in the same way that a child pulls on a plant to make it grow faster. (Havel, 1992).

Identical comments would apply to the utopian mindset promoted by the founders of the Planning, Programming, and Budgeting System (PPBS) in the 1960s, for instance, and more broadly to the "true believers" who consider that indicators are the vectors of absolute rationality.

As already highlighted, "mechanical objectivity is most valued when decision-making is dispersed, when it incorporates diverse groups, when powerful outsiders must be accounted to, when decisions are public and politicized, and when decision-makers are distrusted" (Espeland and Stevens, 1998: 331). The fact remains that *indicators may be either the worst or the best of things depending on whether their use is "mechanical" or "political". This holds true for both the steering of organisations and the steering of policies and their evaluation. In other words, the virtues of referring to the "mechanical objectivity" of indicators do not justify their "mechanical use"*. Indicator users sometimes appear to be deploying the same relentless efforts as the child that Havel refers to – pulling on the plant to make it grow, believing that the objectivity of indicators combined with the actors' absolute rationality will suffice to successfully complete the job.

Today policy makers and societies would run major risks were they to endorse or even tolerate this mindset. The danger is even greater given the boom in indicator-based tools at all levels – micro to macro – of today's world. Moreover, they have now spread across national borders to become sorts of global standards and benchmarks. Over the past fifty years, the massive development and use of indicators has been recommended and fuelled by the initiatives

of public and private supranational organisations (OECD, WB, IMF, ISO, rating agencies, etc.). Certainly, a good number of indicators had already been deployed to measure efficiency or effectiveness for *ad hoc* purposes in specific national or local contexts. The supranational institutions, however, have selected them for other purposes and applied them to global contexts. Indicators are now intended to compare and benchmark national situations, but in the main these national arenas have not been involved as fully fledged stakeholders. With a panoply of scoreboards measuring economic outputs, wellbeing, education, public safety, political stability, etc. as well as standards, certifications and rankings, the increasing importance of indicators veils over local orders (March, 1962) and sidelines entire swathes of reality experienced by different actors – all for sake of having tools that will enable a *relative estimation* of value through comparison.

Assuming that contexts do not really matter and need not be taken in account, naturalised measures such as indicators build and legitimate a kind of soft power that impacts resources and influences entities in many ways – access to financial market funding, foreign investments, interest rates, attractiveness of public schools, displacement of international student flows, etc. The role played by rating agencies during the 2008 global financial crisis, the obsession for international rankings in university reform policies worldwide, the influence of the PISA survey on the perceived quality of secondary education in a number of countries, and the allocation of IMF or World Bank funds to ailing South American or African countries conditioned to structural reforms monitored by performance indicators are among the many examples that evidence the overwhelming influence this soft power has gained in the name of efficiency and effectiveness.

In our societies, where there is a strong temptation to run policymaking as one runs a machine[30] (Foucault, 1980) quite simply because "it is practical", credence is often given to indicators quite simply because they exist and are thus presumed to convey an objective vision of the reality they claim to measure. The problem is that indicators list value criteria that are often distant and arbitrary, sometimes multiple and conflicting. The visions of the world they portray are fuelling growing dissent as our faith in Science and the State wanes. Hence, indicators are less and less seen as embodying authority and as a legitimate expression of public good.[31] Used mechanically, *indicators impose the strength and sometimes the violence of their soft power through rewards and sanctions. This raises a broader question. In what conditions will indicator-based government technologies be perceived as locally relevant and legitimate? This question is about the "sense of measurement", which has a*

[30] «Studying the manner in which one has sought to rationalise power ... showing the important role that has been played by the theme of the machine, perspective, supervision, transparency, etc., neither means that power is a machine, nor that such an idea came about mechanically!" (Foucault, 1980. Author's translation)

[31] For example, the French population are no longer willing to unquestioningly accept, as it did in 1986, the joint declarations of the Reason of Science and the Reason of State certifying that the level of radioactivity measured on the French territory by the "competent authorities" was not impacted by the Chernobyl cloud.

threefold meaning: "the signification embodied in measurement", "the reflexive usage of measurement" and «the sense of proportion» it displays.

This question first raises the issue of the integrity of measurement. The European Commission has, for example, entered the public debate on over-fishing and the depletion of fisheries resources by imposing fishing quotas on bluefin tuna. In this instance, the fishermen are at the same time judge and jury. With no special maritime police to enforce the rules, it is extremely difficult to give credence to the declarations of volumes fished or compliance with fishing limits. *To be considered as reliable, an indicator must be guaranteed by a reliable and independent production chain backed by quasi-legal external supervision that protects it from outside political pressures.*

A second point relates to the way public or private authorities can play around with indicators for communication purposes. Hong Kong for instance lives with an "extremely high» level of pollution 20% of the year, "high" 70% of the year and "moderate" for the remaining 10%. A frightening account, which becomes even more so when we discover that this index is "incredibly tolerant ... it systematically informs the general public that the levels of exposure considered to be dangerous by the WHO are acceptable".[32] *For an indicator to be reliable, we need to have a reflexive stance that helps us to uncover the facts behind the labels in order to counteract possible rhetorical manipulations.*

A third point connects to the way that communities of practice pay attention and give meaning to specific indicators. Members of a given community know how to distinguish the important indicators from those that matter much less and are even discarded (for an example, see Crague, 2005). Some are considered relevant to steering an organisation, while others are used merely as vectors of discourse. This is why companies sometimes produce indicators as a matter of form – for instance, in response to a government demand – and no one attaches the slightest importance to them, as everyone well knows how to identify the truly relevant signals for action-taking. A effective way of undermining the credibility of an indicator is to announce its function but exclude it as a decision-making criterion. *The importance of an indicator depends on the attention its audiences pay to it. This attention is directly related to the effective consequences that the indicator may generate for its audience.*

A measurement may be technically valid or invalid. However, there is no "right measurement" in absolute terms that would justify its being used mechanically. Appropriate use implies dismissing what the efficiency of organisational action and the effectiveness of public action owe to complex power systems with their conflicts, their alliances, their allegiances, their skewed information, their varying degrees of willingness to be informed, etc. (Spenlehauer, 2011). Far from the technocratic, ballistic, planning-oriented and

[32] *Le Monde* newspaper, 27 January 2012, «Planète» feature. (Author's translation)

apolitical use of indicators, a sense of balance argues that their accuracy and validity stem mainly from the strength of the shared conventions underpinning their objectivity and legitimacy: the firmer and more widely shared the indicators, the more they are recognised as being a fair representation of reality. As such, they can provide stronger buffers against vested interests, lobbies and politicking. Not only can they vouch for the impersonal commitments made in their name and on which their credibility as tools of government depends, but they can also help to create contexts that reinforce joint learning capacities. Indicators are not only tools for control and sanction, but also for sharing diagnostics. They take on their full meaning and usefulness when those being evaluated and those evaluating them take up ownership of these two functions simultaneously.

In this respect, evidence repeatedly shows that the major risk is for the diagnostic function to be reduced to no more than a control function. The temptation exists, for example, to characterise an entity solely on the basis of a synthetic number, grade or rank constructed as the weighted sum of the scores it collects along a set of diversified indicators. The prevailing hope is that the merits of this entity can be assessed and positioned mechanically on a unidimensional scale, as can be seen in the current profusion of rankings claiming to provide an easy guide to resolving uncertainty in all areas of existence, be it in the choice of a wine, hotel, university, investment-friendly country or film. The illusion here it that it is both possible and appropriate to avoid exploring the entity's cardinal strengths and weaknesses as if, in real life, an object or the performance of a social group in a given field can be classified as either "good" or "bad".

Two final recommendations could thus be made.

Firstly, it is important that the allocation of public resources should never become the monopoly of a sole "principal". On the contrary, it should develop a diversity of windows and suppliers, as this would allow for greater diversity when the characteristics of evaluated entities and their relative strengths and weaknesses are being scrutinised by the decision-making panels that allocate funds. An additional caveat is that these panels should include members with differentiated stakes and who reflect diversity in terms of gender, educational background, country of origin, etc.

The second important recommendation is that the process and criteria for resource allocation should not be based solely on a mechanistic evaluation of performance[33] *(cf. for example, Gingras, 2008). Indicators are helpful supplementary tools that can be mobilised for policymaking and management purposes.* They provide various proxies that characterise an object, individual or group in its diversity, the aim being to better inform decision-makers so that they become aware and anticipate the implications their actions will have for the different stakeholders. This, for example, is what the current development of mapping

[33] ... moreover, often requiring a large number of audit mechanisms, which gives rise to other difficulties that we do not address here.

methods is trying to achieve by describing in increasingly sophisticated ways the differences and heterogeneities of given entities. These methods aim to shed light on their strong and weak points, provide a medium for strategic communication between the evaluators and the evaluated, and enable judgments and resource allocations to be anchored in a shared reflexive approach. This creates fertile ground for inventiveness and originality. They may also dissuade reformers and cost-cutters from over-hastily sacrificing public service missions on the altar of market regulation ideologies. For instance, should treatment for patients be halted or a hospital closed down when their costs are not in line with the volume of treatment specified by performance indicators designed by some remote cost control agency? While the answer is far from simple, raising the question has the merit of forcing some in-depth decisions, rather than simply resorting to the "rituals of verification" to avoid the issue (Power, 1997)... a contemporary version of the old bureaucratic routine so insightfully described by Robert Merton (1940).

References

Altfeld, M.F. and G.J. Miller (1984), "Sources of Bureaucratic Influence: Expertise and Agenda Control", *The Journal of Conflict Resolution*, 28(4): 701–730.

Appadurai, A. (1996), *Modernity at large: cultural dimensions of modernity*, University of Minnesota Press, London and Minneapolis.

Barbier, J.-C. (2011), "L'expérimentation du Revenue du Solidarité Active (RSA) en France (2007-2009)", Cahiers de l'Evaluation.

Boltanski, L. and L. Thevenot (1991), *De la justification, Les économies de la grandeur*, Gallimard, Paris.

Bourguet, M.-N. (1988), *Déchiffrer la France, La statistique départementale à l'époque napoléonienne*, Editions des Archives Contemporaines, Paris.

Clotfelter, C.T. (2011), *Big-time sports in American universities*, Cambridge University Press, Cambridge, MA.

Courpasson, D. (2000), *L'action contrainte : Organisations libérales et domination*, PUF, Coll. *Sciences sociales*, Paris.

Crague, G. (2005), "Le travail industriel hors les murs – Enquête sur les nouvelles figures de l'entreprise", *Réseaux*, 134: 65–89.

Dahl, R.A. (1961), *Who Governs?: Democracy and Power in an American City*, Yale University Press, New Haven.

Delors, J. (ed.) (1971), *Les indicateurs sociaux*, Futuribles, SEDEIS, Paris.

Desrosières, A. and L. Thévenot (2006 [1988]), *Les catégories socio-professionnelles*, La Découverte, *Repères* collection, Paris.

Desrosières, A. (1993), *La politique des grands nombres, Histoire de la raison statistique*, La Découverte, Paris.

Di Maggio, P.J. and W.W. Powell (1983), "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields", *American Sociological Review*, 48: 147–160.

Enthoven, A.C. and K.W. Smith (2005), *How Much is Enough: Shaping the Defense Program 1961-1969*, Rand Corporation, Santa Monica, CA.

Ehrenberg, R.G. (2000), *Tuition Rising, Why College Costs So Much*, Harvard University Press, Harvard.

Espeland, W.N. and M. Sauder (2007), "Rankings and Reactivity: How Public Measures Recreate Social Worlds", *American Journal of Sociology*, 113(1): 1–40.

Espeland, W.N. and M.L. Stevens (1998), "Commensuration as a social process", *Annual Review of Sociology*, 24: 313–343.

Esping-Andersen, G. (1990), *The Three Worlds of Welfare Capitalism*, Polity Press, Cambridge (UK) and Princeton University Press, Princeton NJ.

Ferret, J. and V. Spenlehauer (2008), "Does policing the risk society hold the road risk?", *British Journal of Criminology*, 49(2).

Foucault, M., F. Ewald and D. Defert (eds) (1980), *Dits et écrits 4 1954-88: 1980-88*, Gallimard, Paris.

Fourquet, F. (1980), *Les comptes de la puissance, Histoire de la comptabilité nationale et du plan*, Editions Recherches Encres, Paris.

Frank, R. and P.J. Cook (1995), *The Winner-Take-All Society*, Martin Kessler Books at the Free Press, New York.

Fubini, D., C. Price and M. Zollo (2006), *Mergers: Leadership, performance, and corporate health*, Palgrave Macmillan, New York.

Gibert, P. and M. Andrault (1984), "Contrôler la gestion ou évaluer les politiques", *Politiques et Management Public*, 2: 123–132

Gibert, P. (1985), "Management public, management de la puissance publique", *Politiques et Management Public*, 4(2).

Gibert, P. (2003), "L'évaluation de politique : contrôle externe de la gestion publique?", *Revue Française de Gestion*, 6(147): 259–273.

Gingras, Y. (2008), "La fièvre de l'évaluation de la recherché, Du mauvais usage de faux indicateurs", research paper 2008-05, CIRST.

Granet, M. (1988 [1934]), *La pensée chinoise*, Albin Michel, Paris.

Hatchuel, A. and B. Weil (1995), *Experts in Organization. A Knowledge-Based Perspective on Organizational Change*, Walter de Gruyter, Berlin–New York.

Havel, V. (1992), Speech to the Académie des sciences morales et politiques, Paris, 27 October 1992, quoted in Foreign Broadcast Information Service and Joint Publications Research Service report (1992), "East Europe", NTIS, Virginia, 22 Dec.

Karpik, L. (1996), "Dispositifs de confiance et engagements crédibles", *Sociologie du Travail*, Vol.4.

Karpik, L. (2010), *Valuing the Unique: The Economics of Singularities*, Princeton University Press, Princeton.)

Laufer, R. and C. Paradeise (1982), *Le prince bureaucrate*, Flammarion, Paris.

La sécurité routière en France, 2011, Bilan de l'année 2011, http://www.securite-routiere.gouv.fr/la-securite-routiere/l-observatoire-national-interministeriel-de-la-securite-routiere/bilans-annuels/bilans-annuels-de-la-securite-routiere-en-france

Lindblom, C.A. (1959), "The Science Of 'Muddling Through'", *Public Administration Review*, 19(2): 79–88.

Mallard, G., C. Paradeise and A. Peerbaye (eds) (2009), *Global Science and National Sovereignty, Studies in Historical Sociology of Science*, Routledge Studies in the History of Science, Technology and Medicine, New York–London.

Mandl, U., A. Dierz and F. Ilzkovitz (2008), "The effectiveness and efficiency of public spending", Economic Paper 301, *Economic Papers*, European Commission.

March, J.G. (1962), "The Business Firm as a Political Coalition, Journal of Politics", 24: 662–678.

Meny, Y. and J.-C. Thoenig (1989), *Les politiques publiques,* Presses Universitaires de France, Paris.

Merton, R.K. (1940), "Bureaucratic structure and personality", *Social Forces*, 18: 560–568.

Miller, P. (2001), "Governing by Numbers: Why Calculative Practices Matter", *Social Research*, 68(2): 379–396.

Mintzberg, A. (1979), *The Structuring of Organizations*, Prentice Hall, New Jersey.

Musselin, C. (2001), *La longue marche des universités françaises*, PUF, Paris.

Ocqueteau, F. and J.-C. Thoenig (1997), "Mouvements sociaux et action publique : le transport routier de marchandises", *Sociologie du Travail*, 39(4): 397–424.

Paradeise, C. (2012), "Universités : un marché mondial de la connaissance", *Les Cahiers français*, No. 367.

Perret, B. (2006), "De l'échec de la rationalisation des choix budgétaires (RCB) à la loi organique relative aux lois de finances (LOFL)", *Revue française d'administration publique*, 1: 31–41.

Phélippeau, E. (2003), "Conseil national de l'évaluation, Commissariat général du plan, La politique de sécurité routière, Les systèmes locaux de contrôle-sanction", Report by the evaluation authority presided by Michel Ternier.

Porter, T.M. (1995), *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton University Press, Princeton NJ.

Power, M. (1997), *The AuditSsociety, Rituals of Verification*, Oxford University Press, Oxford.

Ravinet, P. (2011), "La coordination européenne 'à la bolognaise', Réflexions sur l'instrumentation de l'espace européen d'enseignement supérieur", *Revue française de science politique*, 61(1).

Rose, N. (1991), "Governing by numbers: Figuring out democracy", *Accounting Organizations and Society*, 16(7): 673–692.

Schlenker, J.-M. (2009), "Utilité et limites des indicateurs bibliométriques", presentation at the Alembert seminar, University of Paris Sud, 24-05, Orsay.

Spenlehauer, V. (2011), "Des sciences sociales engagées, L'évaluation des politiques publiques", thesis for the habilitation to direct research, University of Paris Est.

Sperber, M. (2000), *Beer and circus. How big-time college sports is crippling undergraduate education*, Holt paperback, Henry Holt and Cie., New York.

Stark, D. (2009), *The sense of dissonance, Accounts of Worth in Economic Life*, Princeton University Press, Princeton.

Thoenig, J.C. (1971), "Le PPBS et l'administration publique, Au-delà du changement technique", *Annuaire International de la Fonction Publique,* pp.97–114, Institut international d'administration publique, Paris.

Weick, K. (1976), "Educational organizations as loosely coupled systems", *Administrative Science Quarterly*, 21: 1–19.

Wildavsky, A. (1969), "Rescuing Policy Analysis from PPBS", *Public Administration Review*, 29(2): 189–202.

Conférences & Séminaires

Part 4: Applying Evaluation to Development and Development Aid

5. Dimensioning Development Aid: Some Lessons from Evaluation

Ruerd Ruben, Centre for International Development Issues (CIDIN), Radboud University Nijmegen; Ministry of Foreign Affairs, the Netherlands

Abstract

Most discussions on the development effectiveness of aid tend to focus on aggregate flows and neglect the various categories of aid and different aid delivery mechanisms. We emphasise, however, that the key question is not whether aid works, but which aid works. The potential effects – both beneficial and adverse – of aid appear to be mainly the consequence of how aid is provided. This paper therefore addresses two questions: which aid works and how aid is delivered. These issues are becoming increasingly relevant given the declining societal trust in public aid and against the background of the growing interests in the new aid architecture.

We discuss three basic, albeit usually underestimated aspects that critically influence development effectiveness: (a) resource complementarities between different programme components, (b) substitution effects between different activities, and (c) spillover effects that influence aid effectiveness at aggregate level. We present some empirical examples of these mechanisms and indicate their particular relevance for the new types of institutional arrangements that characterise upcoming reforms of the international aid architecture (public-private partnerships, multi-donor trust funds and civil society support funds).

5.1. Introduction: heterogeneity of aid

For a long time, development aid received unconditional citizen support in almost all European countries. The latest Eurobarometer report (EC, 2010) registers, however, that the share of people considering aid as "very important" to help people in developing countries has declined from 53% in 2004 to 45% in 2010. At the same time, it is striking that most citizens largely overestimate the volume of aid provided by their government.[34] People also seem more reluctant to support professional development agencies and prefer to practically engage in small-scale development cooperation activities.

The societal support base for development aid is only partly influenced by information on or conceptions about aid effectiveness. Overestimation of the dimensions of aid easily leads to unrealistic expectations. Moreover, general scepticism about the role of (public and private) institutions in managing socio-economic crises is reflected in a gradual shift from "trust me" and "tell me" to "show me" (and sometimes also "involve me") attitudes. This is further exacerbated by the rather unfruitful dialogue amongst development professionals in which extreme views may be voiced, ranging from overoptimistic ideas that massive aid can eradicate poverty (Sacks) to totally pessimistic views that most aid is wasted (Easterly) and perpetuates dependency (Moyo).

Few of these perceptions and debates are based on detailed empirical analyses of aid effectiveness. These studies thus usually suffer from two major shortcomings: *(a)* little attention is given to the "framing" of aid as a – usually minor – component of the overall resource flows dedicated to development; and *(b)* most aid programmes are almost exclusively perceived from the supply side (donor perspective) and tend to disregard demand-side criteria regarding the tailoring of aid to local requirements and preferences. Recognising these dimensions enables us to consider the real importance of aid and to outline possible pathways to enhance development effectiveness within a wider framework of a renewed (inter)national aid architecture.

A wide range of studies has specifically addressed the question of "whether aid works", focussing on cross-country and panel data evidence derived from the relationship between aid and growth at macro-level. The results from the ongoing debate are rather inconclusive, and lead – given the heterogeneity of aid and the relatively small aid volume compared to

[34] Data from the World Opinion Poll indicate that Americans guessed that the US spends 25% of the budget on foreign aid, but opined that the figure should be about 10%. The actual US number is 0.21% of GDP.

other financial flows – to conclusions that effects are rather limited, highly context-dependent and only visible in the medium to long run (Arndt *et al.*, 2010). On the other hand, the empirical literature has devoted much attention to the determinants of the geographical and sector allocation of aid. It is assumed that donors are mainly concerned with two questions: "To whom should we give aid?" and "How much aid should we provide?" Hence, the majority of studies make the implicit assumption that all donors give similar types of aid and use the same channels. It can be argued, however, that a donor's choice set is far more diverse (Raschky and Schwindt, 2011; Lessmann and Markwardt, 2010). Donor countries do not only have to make a decision on the amount and recipients of aid but also on the preferred transfer channel (bilateral, multilateral or civic aid) and the type of delivery modes for aid (cash or kind, conditional transfers, loans or grants, etc.).

In this paper we outline some considerations regarding the effectiveness of different types of development programmes through the prism of the ways in which aid is organised as a component of the funding of multifaceted and often intertwined development efforts. Shifting attention from "aid effectiveness" to "development effectiveness" means that we need to identify the complex interplay between development aid and local efforts, and discuss the incentive structures that make it possible to adequately dovetail different types of foreign aid with local resources. We emphasise, therefore, that the key question is not *whether* aid works, but *which* aid works. The potential – beneficial and adverse – effects of aid appear to be mainly the consequence of *how* aid is given (Barker, 2011; Bourguignon and Sundberg, 2007). This implies that different types of aid providers may offer specific incentives to individual clients. The rationale for selectivity can thus be based on particular (dis)advantages of aid delivery procedures for reaching clients in specific settings .

Based on several programme evaluations and impact studies conducted by the Policy and Operations Evaluation Department (*Inspectie Ontwikkelingssamenwerking en Beleidsevaluatie* – IOB) of the Netherlands' Ministry of Foreign Affairs, among others, we outline three important, albeit usually underestimated, aspects that critically influence development effectiveness: *(a)* resource complementarities between different programme components, *(b)* substitution effects between different activities, and *(c)* spillover effects that influence aid effectiveness at aggregate level. We present some empirical examples of these mechanisms and indicate their particular relevance for the new types of institutional arrangements that characterise the upcoming reforms in the international aid architecture (public-private partnerships, post-conflict reconstruction programmes and civil society organisations).

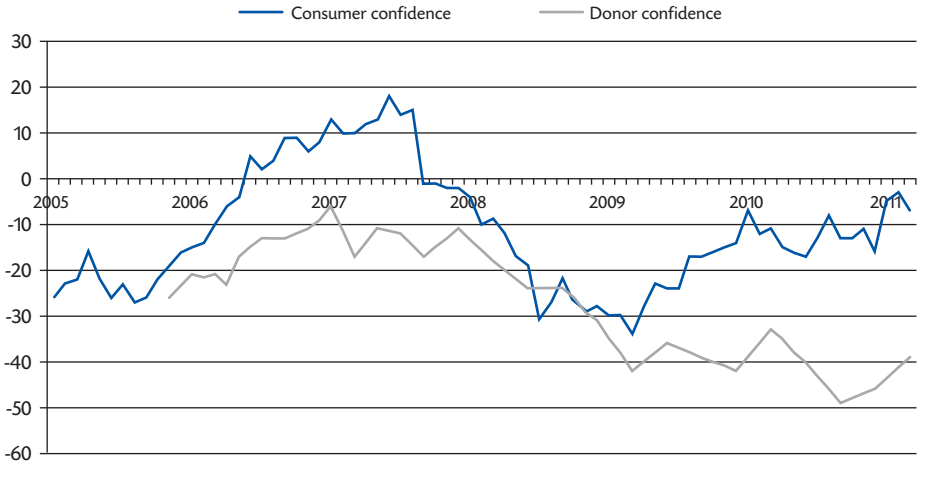
5. 2. Development cooperation between trust and scepticism

Development cooperation suffers from declining public support; in many OECD countries, the reduction of aid budgets is mentioned as a device for reducing public spending.[35] In The Netherlands, confidence in charities is declining substantially faster than general consumer confidence (see Figure 3), indicating a more than proportional reduction in donor trust.

The current literature largely fails to throw significant light on the key drivers of the declining public support for development assistance across the population in many donor countries (a notable exception is found in van Hudson and Heerde, 2010). It has been suggested that three factors are of key importance: *(a)* credible evidence of successes and failures, based on systematic

Figure 3

Consumer and donor confidence (the Netherlands, 2005-2011)



Source:Dutch Donations Panel (December 2011), www.wwav.nl

[35] Results from the tracking surveys undertaken by DfID confirm that there is an established downward trend in public support for increased action by the UK government towards reducing poverty in developing countries. Thus, only 35% of respondents supported increased government action in February 2010, compared to 50% in September 2007. While 55% of respondents were of the view that the government should spend more on aid to developing countries in September 2007, this support had declined to 40% by February 2010.

evaluation of outcomes and impacts of development programmes, *(b)* information about the governance and delivery of development aid and local institutional implications, and (c) opportunities for direct citizen involvement in aid programmes.[36]

Several interrelated problems may be responsible for the declining trust in international aid. First, on the supply side, the proliferation of aid agencies leads to the fragmentation (i.e. aid spread over many projects, programmes or sectors) and possible duplication of efforts (Koch, 2009; Schulpen *et al.*, 2011). Second, on the demand side, the effective use of aid is challenged by risks of undermining local governance and increasing corruption (Frot and Santiso, 2008). Third, at the interface between supply and demand, the wide diversity of modalities for contracting and delivering aid leads to high transaction costs and a dearth of incentives intended to direct capacity development towards results-based aid management (Gibson *et al.*, 2005).

Funding for international cooperation is nowadays disbursed from a wide diversity of sources, ranging from ODA contributions paid from public budgets (taxes) to money raised from the general public (donations to charities). Along the public-private continuum, several new organisations have emerged, ranging from large-scale philanthropic foundations (Gates, Rockefeller, Ford) to numerous small-scale private initiatives. Entrepreneurial co-funding (e.g. Public-Private Partnerships or PPPs), development operations made possible by equity funds (e.g. the companies UNITUS and Elevar that offer market-based solutions to poverty), stock operations (e.g. the emission of vaccine bonds) based on aid pledges (e.g. IFFIm-GAVI), and diaspora bonds (tapping remittance flows from migrants) have created innovative procedures for development funding. Consequently, development finance has become far more diversified thanks to the blending of donations, lending and borrowing, bond finance and future-flow securitisation (Ketkar and Ratha, 2009).[37]

There is growing consensus that the current proliferation in the international aid architecture has resulted in structures and institutions for supplying development assistance that are not fit for the purpose. New institutions are often created, but old ones are almost never shut down. As a result, there are now more than 1,000 mechanisms for supplying development finance. Similarly, there has been a proliferation of global partnerships and initiatives for multi-donor trust funds (MDTFs) in recent years, generally customised to mobilise finance for a single issue. There are already more than 100 disease-specific global partnerships active in

[36] A more detailed analysis (see Kinsbergen *et al.*, 2011) indicates that volunteers are somewhat sceptical towards established development organisations, but crowding-out is relatively limited. Corroborating the proximity hypothesis, volunteers who perceive a smaller distance to beneficiaries spend more volunteering hours in private development initiatives (PDIs).

[37] According to a Hudson Institute Center for Global Prosperity report (Center for Global Prosperity, 2010), total official development assistance was USD 120 bn in 2009, while global philanthropy was USD 53 bn. Private capital investment (USD 228bn) forms the largest financial flow from richer to poorer countries, while remittances (USD 174bn) were the second largest flow.

the field of health alone.[38] Many global initiatives use a "vertical" programming approach, implementing a standard set of programmes in a specific sub-sector across all countries of operation. Scepticism towards MDTFs is mounting and aid financing is sometimes considered as a major impediment to effective poverty alleviation (Barakat, 2009). At times, funding approaches are not consistent with either the principles of donor harmonisation or the alignment with country strategies and systems.

There is a growing recognition among donors that their core business is contributing to broader development effectiveness, not just aid effectiveness (Kindornay, 2011). The current policy challenge is to enhance development effectiveness through *(a)* greater impact of development programmes on achieving human development and improving the lives of the poor, and *(b)* strengthening policy coherence between aid agencies and across support areas, as well as *(c)* improving the organisation and governance structures of aid (including predictability and timely availability). Development effectiveness thus involves paying attention to delivering results at different scales ("Better Aid") embedded within an institutional framework of international partnership relationships ("Better Aid Architecture").

In this context, the organisation of aid has become in itself a key factor for enhancing public trust with respect to the achievement of development effectiveness. Providing better insights into the underlying mechanisms that influence development effectiveness might be of critical importance in re-establishing societal support for international aid. The role of aid in its specific context thus deserves far more attention. We therefore discuss three somewhat neglected institutional conditions that are of fundamental importance to understanding the development effectiveness of aid programmes.

[38] Global partnerships have been quite effective as vehicles for resource mobilisation, accounting for 3% of ODA in 2005. However, there are widespread concerns, especially in the health sector, that global partnerships are diverting resources away from the development of national health systems.

5.3. Searching for aid complementarities

Aid effectiveness is usually related to specific interventions and involves trying to identify relationships between aid inputs and development outcomes. This analysis is already complicated due to the fact that international aid is only one of the many components that contribute to the final result. These outcomes should be conceived as resulting from multiple – often heterogeneous – contributions that could potentially reinforce (but also contradict) each other. Optimal sequencing and right proportionalities critically influence aid effectiveness.

The concept of complementarities has its origin in contingency theory. Milgrom and Roberts (1995) proposed that some organisational activities and practices are mutually complementary and tend to be adopted together, each enhancing the contribution of the other. Consequently, the impact of a system of complementary practices will be greater than the sum of its parts because of the synergistic effects of bundling practices together. Complementarities thus indicate a condition of increasing returns in which adopting (doing more of) an activity (e.g. implementation of certain aid strategy) has a higher pay-off when simultaneously adopting (doing more of) a complementary activity (e.g. implementation of another aid strategy).[39]

Complementarity thinking has its basis in classical growth models (assuming strict complementarity between capital and labour), but is more generally applied in the analysis of options for reinforcing enterprise-level efficiency gains. Sarker *et al*. (2001) extend the framework to multi-agency alliances where several partners contribute to resource complementarities. Their analysis reveals that partnership characteristics indirectly affect performance through certain mediating behavioural variables (i.e. trust, reciprocity, communication). The socio-psychological aspects embodied in relationship capital are important since they act as coordinating mechanisms and determine the quality of the relationship in the collaboration. Boyer (2005) further outlines some critical conditions for reaching institutional complementarities, related to resource heterogeneity, differences in information and possibilities for overcoming binding constraints.

Recent applications of complementarity thinking to international aid address both the optimal *modality mix* (e.g. combining multilateral, bilateral and civic funding; mixing loans and grants, combining financial and technical assistance, etc.) and the most suitable alignment of the *activity mix* (e.g. combining international and local resources).

[39] Complementarity theory essentially follows contingency theory, which considers performance as dependent on "fit" – generally investigated in terms of moderation – between financial and organisational variables.

While the former issue refers mainly to supply-side motives for reinforcing aid complementarities and reducing fragmentation at aggregate level, the latter also includes attention for demand-side criteria regarding the contribution of aid to improved factor mobilisation and productivity (Foster and Leavy, 2001).

Aid complementarities receive major attention in international debates regarding donor coordination (e.g. Paris Agenda; Busan HLF-4, etc.). Aid harmonisation and alignment of donor practices are generally considered to be key factors for enhancing aid effectiveness. Far less attention is given to complementarities at more concrete levels, as for example, the preferred combination of different types of funding (project aid, programme aid, sector support, budget support, etc.), the degree of concessionality (mixture of loans and grants) and the selection of the most appropriate aid delivery channels (bilateral, multilateral, civic or private). Equally important are complementarities between different kinds of interventions (e.g. in education: teachers and books, classrooms and teaching staff; in health care: health workers and training). Primary motives for the choice of specific aid modalities by donors ought to be based on criteria of comparative advantage and transaction costs. At a more political level, however, aspects of public administration, public sector accountability and national resource mobilisation (tax revenues) also play a role in determining aid management capacities and prospects for national ownership (see, Ohno and Niiya, 2004).

The likelihood of achieving development impacts tends to be related to the *right combination* of development efforts, both from external donors and local partners. This refers to the possibilities for creating *synergies* between different (public and private) programmes that are able to overcome critical poverty thresholds at client, community or regional level. On the other hand, individual clients may simultaneously receive similar services through different providers, and there may be considerable overlaps in aid allocation that produce declining marginal results. The multiple and sometimes overlapping support modalities may reinforce each other (usually referred to as "complementarities") but can also compete for the same resources and capacities ("trade-offs").

Mavrotas (2005) uses disaggregated data on aid modalities (project aid, programme aid, technical assistance and food aid) for Uganda during the 1980–1999 period to test the differentiated impact on fiscal response (e.g. public investment and consumption). Using reduced-form equations, he obtains results indicating that project aid and food aid lead to a reduction in public investment, whereas programme aid and technical assistance are positively associated with changes in public investments. Further analysis regarding the sources of each type of aid (most bi/multilateral aid for programmes; NGO aid more oriented towards projects and capacity building) could provide an indication of the effects of choosing a specific channel.

Lessmann and Markwardt (2010) consider different aid modalities – together with measures of political and fiscal decentralisation – in a classical growth model to identify differences in aid effectiveness. They distinguish five types of public aid (i.e.

grants, loans, technical assistance, humanitarian aid and total net ODA) and also differentiate between the various sources of aid (bilateral vs. multilateral). Aid effectiveness is considered to depend on the *interaction* between external aid delivery and internal policy conditions. Results indicate that loans have more impact than grants on public sector decentralisation, while decentralisation enhances the growth pay-off of technical assistance (but the inverse is true for humanitarian assistance), and bilateral aid is slightly more growth-enhancing than multilateral aid.

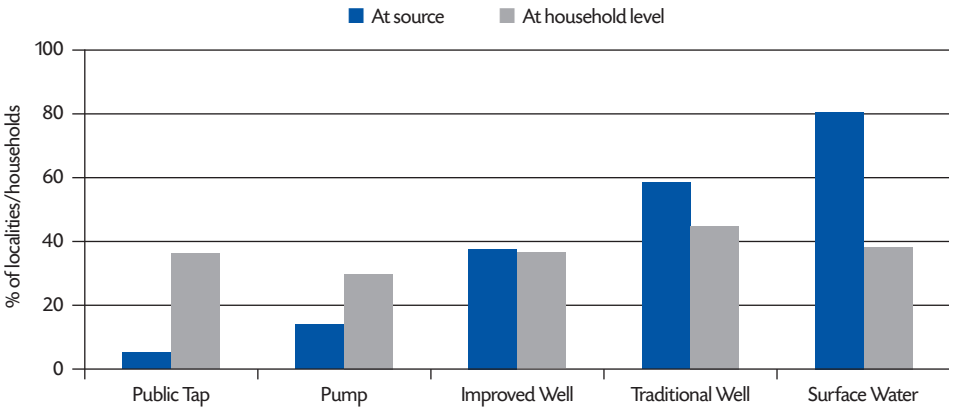
In a similar vein, Raschky and Schwindt (2011) take a large number of donors and analyse the reasons behind their choice of either bilateral or multilateral channels to allocate disaster assistance. Disaster aid is therefore differentiated in cash or in-kind deliveries. Using seven years of OCHA data, results suggest that the donor's choice of the delivery channel and type of aid is mainly driven by the quality of institutions in the recipient country (multilateral aid is preferred for more remote countries with severe rule-of-law problems), as well as by strategic and natural resource interests (more bilateral aid for relevant trading partners and the aligned UN voting pattern)

The fact that aid complementarities matter for development effectiveness can also be illustrated with two typical examples derived from recent programme evaluations. IOB conducted a rigorous impact evaluation of rural water supply and sanitation programmes in Benin covering several interventions to which a number of donors have contributed (IOB-BMZ, 2011). The objective of the support to water supply and sanitary facilities goes beyond sustainable access: it aims to reduce the burden of water collection (typically a task for women and girls), improve health, raise school enrolment and attendance, improve livelihoods and, ultimately, reduce poverty. The study seeks to determine whether these effects materialise. Special attention is given to comparing the water quality from different sources and to identifying the linkages between water use, sanitation and hygiene behaviour. The importance of sanitation is illustrated through the incidence of E.coli at different water sources. Interestingly, public taps from the project deliver far cleaner water at source, but – due to insufficient sanitation measures – a great deal of this benefit is lost at user level (see Figure 4). Water contamination due to transport and storage cannot be addressed only through technical measures (water tabs provided by public agencies) but requires thorough attention for training and consciousness-raising (usually provided by local NGOs).

These complementarities between drinking water and sanitation measures deserve due attention and might also require a combination of financial and technical support through a mixture of public and civic aid delivery modalities. The widely assumed comparative advantage of NGOs in supporting local training, capacity building and ownership can be used to enhance the linkages between hardware and software programme components. The drawback of this alliance might be that NGO aid becomes strongly clustered around bilateral programmes, thus reinforcing the spatial skewness in aid distribution and weakening the poverty targets of NGO aid (Koch, 2009; Fruttero and Guari, 2005).

Figure 4

E-coli incidence at different water sources (Benin)



Source: IOB-BMZ (2011)

In the field of agricultural development and food security, similar complementarities are found in programmes focussing on land titling and registration. Fort (2008) draws on the experience of evaluating the national Titling and Registration Program in Peru (PETT) to present fresh evidence on the impact of this type of programme. This study concludes that land titling and registration could enhance a tenant's landholding security, but that complementary policies are needed to materialise the potential effects. Providing land titles clearly increases the local farmer' subjective sense of ownership and may increase demand for credit, but if an additional supply of rural credit is made available, this only results in more on-farm investment. The critical complementarity between titling and credit provision is frequently overlooked and the net results of land titling programmes may therefore remain rather disappointing. Similar complementarities are also found in programmes to enhance land productivity (requiring both technical and financial support) and in programmes focussing on value chain development (based on business-to-business linkages with additional NGO support to strengthen farmers' organisations).

The general lessons than can be drawn from these evaluation studies indicate that aid effectiveness might be strongly enhanced if a *conscious combination* of interventions is pursued, sometimes also involving different donor agencies (that offer specific types of aid). This requires a thorough analysis of the most limiting factors and critical constraints to poverty reduction in each specific context. Moreover, the particular *mix and sequence* of aid modalities could also be helpful in developing pathways that gradually

reduce aid dependency by changing the composition of external and local contributions to development programmes. However, aid complementarities should not be managed from the supply side alone, but also require careful management on the demand side in order to verify whether the incentives intrinsic to the aid delivery regimes are consistent with the behavioural motivations of the receiving agents.

5.4. Identifying substitution effects

Substitution effects occur when domestic resources are switched away from activities supported by foreign aid, thus changing the availability of resources for other remaining activities. In analogy with a price change, aid recipients are likely to replace expensive domestic activities with less costly – or more rewarding – alternatives supported by foreign aid. Otherwise, the aid fungibility literature suggests that local external project funds are preferably used for the purchase of tradeables (investments), whereas domestic resources are allocated to non-tradeables and the financing of recurrent costs (Feyzioglu *et al*., 1998).

The general expectation that aid enhance macroeconomic growth through domestic savings and investments has proved difficult to confirm empirically. This is partly due to unrealistically high expectations regarding the impact of aid, but it can also be explained by numerous methodological problems (finding accurate model specifications) and conceptual constraints (selecting adequate growth models).[40] The explanation with the longest history is that aid goes to consumption, thus crowding out domestic savings and investment.

There is a longstanding debate on the implications of international aid for domestic savings and investments (White, 1993). This is related to the fact that "cheap" aid may replace more expensive domestic capital. The net effect depends on the impact of aid on domestic investments and is informed by the implications for the interest rate (usually a negative effect) and for total income (an expected positive effect). In a situation of large deficits, aid has only a minor effect on capital costs but a likely stronger impact on income, and crowding-in may in fact then occur. The crowding-out effect is likely to dominate in the long run only when the economy is operating near full employment.

The academic literature is still somewhat inconclusive about the crowding-in/out effects of aid at aggregate level. Shields (2007) examines the relationship between foreign aid and domestic savings using data for 119 countries. Regressions for each country are run separately in order to identify which countries have a positive aid-saving experience. Countries are categorised according to the strength of the aid-saving relationship. Few countries show evidence of a substantial crowding-out and these findings indicate that aid clearly benefits savings and, hence, stimulated domestic investment.

Masud and Yoncheva (2005) tested whether foreign aid reduces government efforts geared to achieving developmental goals,

[40] There is a vast literature based on endogenous growth models suggesting that in the long run the saving (and investment) rate is less important for growth. Institutional factors influence resource efficiency and the ability of an economy to innovate and respond to opportunities. Numerous growth models have been formulated to estimate the impact of policies and institutions on growth. However, Knack (2001) argues that aid may increase the rewards to rent-seeking behaviour and, hence, undermine the quality of governance.

distinguishing between the different types of aid. To assess aid effectiveness, they compare official bilateral aid with (European) NGO aid flows, using specific human development indicators of poverty reduction (infant mortality and illiteracy). Analysing panel data from 76 countries over 12 years, they find that NGO aid significantly reduces infant mortality (i.e. is more focused on poorer countries and better at reaching grassroots), while bilateral aid registers no significant effect (possibly due to a lack of additionality). For literacy reduction, only government expenditures – with budget support – show significant effects. Moreover, aid through NGO channels does not seem to crowd out public expenditures. There is evidence of a substitution effect between bilateral aid and public social sector expenditures, whereas NGO aid does not generally affect social spending in the recipient country. Similar differentiated effects have been registered for grants compared to loans (Lessmann and Markwardt, 2010) and for multilateral aid compared to bilateral aid.[41] A follow-up paper (Nancy and Yontcheva, 2006) analysing the determinants of aid allocation by NGOs shows that NGOs are more likely to intervene in poor countries with low life expectancy.[42] This is further analysed by Nunnenkamp and Öhler (2009) for different channels of German aid, identifying specific drivers for bilateral aid (strong needs orientation) and NGO channels (more merit-oriented focussing on voice and accountability).

In a recent paper, Arndt *et al*. (2010) re-assess recent contributions to the aid-growth literature – taking inspiration from the programme evaluation literature – to develop a new counterfactual framework that leads to more robust conclusions. The average treatment effect of aid on long-run growth is found to be small but positive and statistically significant with an elasticity of GDP growth to foreign aid of around 0.10, which materialises with a considerable time lag. These estimates are consistent with the view that foreign aid stimulates aggregate investment and may also contribute to productivity growth, even when some fraction of the aid is allocated to consumption.

While most of the empirical analyses have focussed on aggregate macroeconomic level, programme and project evaluations find considerable evidence that, at sectoral and local level, aid may lead to relatively strong substitution effects. In the health sector, there is considerable debate on whether HIV/AIDS control is crowding out other health initiatives, such as vaccination programmes (Schiffman, 2006). Lu *et al*. (2010) show that for every dollar of international health aid provided to governments, government health funding falls by USD 0.43–1.14. Interestingly, development assistance for health to the non-governmental sector had a positive and significant effect on domestic government health spending. These results were robust

[41] Morissey (1990) uses an input-output analysis to estimate and compare the impact of multilateral and tied bilateral aid on the UK economy in 1980 and 1985. The results suggest, quite strongly, that multilateral aid generates greater benefits both in volume terms and per equivalent amount of aid expenditure.

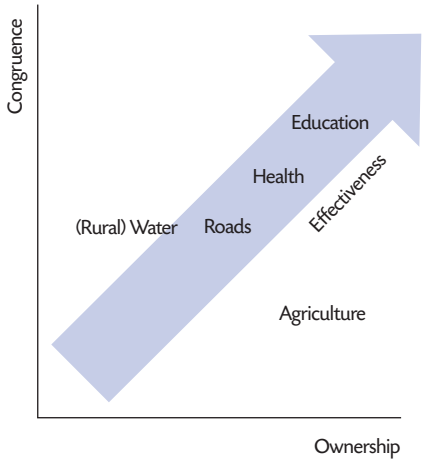
[42] Koch (2009) finds, however, that Dutch NGOs tend to cluster activities around bilateral aid and concentrate their efforts in better-endowed regions.

to different model specifications and subset analyses.

Substitution effects are most widely debated in the area of budget support, usually on the basis of a rather narrow concept of conditionality. In practice, the occurrence of substitution effects is highly dependent on the degree of ownership and the congruence of national policy (see Figure 5).

Figure 5

World Energy Consumption by Source



Budget support in Zambia turned out to be effective in those areas where the support is based on national ownership (e.g. in education and health care, and to a lesser extent in agriculture, where donors and the government disagree about funding priorities). Aid effectiveness is better guaranteed if complementary national budgets are devoted to the sector (as in education and health). If these funds are not available (as in the water sector) project-oriented support is likely to offer better results (de Kemp *et al*., 2011).[43]

In another area, IOB evaluation studies on the impact of sectoral budget and programme support to basic education suggest that there is little evidence for substitution effects (de Kemp *et al*., 2011; IOB, 2011*a*). Free or subsidised basic education in Kenya resulted in fewer children attending school, as private schools for the poor closed down owing to a crowding-out effect induced by the introduction of "free" tuition fees. The net impact could have been at best a simple transferral of children from the private sector to the government sector rather than a net increase in enrolment (Tooley *et al*., 2008; Vos *et al*., 2004). In Zambia, budgetary decentralisation shifted spending from the province to the districts, negatively affecting the equity of fund allocation and crowding out parental and community contributions. In several countries, decentralisation led to lower national government spending, a decline in aggregate education expenditures and a rise in formal and informal payments by parents (Das *et al*., 2004). Lower levels of government were handed over responsibilities without commensurate funding. As a result, teacher salaries and complementary inputs declined and drop-out rates began to rise in some countries (Vandycke, 2000).

Research using randomised control trials (RCTs) conducted within the Mexican PROGRESA programme – in which beneficiary households receive a subsidy

[43] Similar results are found by T. Cordella and G. Dell'Ariccia (2003).

conditional on school attendance – indicates that the programme does crowd out private transfers (Albarran and Attanasio, 2002). The likelihood of receiving private transfers and the amounts received are significantly and negatively affected by the programme. The transfers received from friends and relatives suffer stronger crowding out effects compared to the amounts received from migrant family members.

Ricker-Gilbert *et al*. (2011) use a double-hurdle model with panel data from Malawi to investigate how fertiliser subsidies affect farmer demand for commercial fertiliser. While controlling for potential endogeneity caused by the non-random targeting of fertiliser subsidy recipients, results show that on average one additional kilogram of subsidised fertiliser crowds out 0.22 kg of commercial fertiliser, but crowding-out is much lower among the poorest farmers and higher among non-poor farmers. This indicates that targeting fertiliser subsidies at the rural poor is likely to maximise the subsidy programme's contribution to total fertiliser use.

Raschky and Schwindt (2008) discuss the impact of foreign aid in case of catastrophic events (earthquakes; cholera epidemics) on the level of mitigative activities in aid-receiving countries. They show that the anticipation of foreign aid partly crowds out preventive collective action for *ex-ante* risk management. The crowding-out effect may result in both a lower probability of surviving a disaster and an increase in the magnitude of event-related fatalities or epidemics. Estimates suggest that foreign aid in previous years might crowd out *ex-ante* risk management activities in recipient countries.

Food assistance is one of the most debated forms of support and meets with widespread scepticism regarding its possible influence on local disincentives to work and on the crowding-out of private transfers. There is an abundant literature discussing different aspects of food assistance, including (and certainly not limited to) the incentive effects of such transfers on labour supply (Abdulai *et al.*, 2005), changes in local production through price effects (Tadesse and Shively, 2009), the crowding-out of informal assistance (Dercon and Krishnan, 2003), the effects on productivity due to an improved nutritional status, the effects on asset accumulation in order to break out of poverty traps (Gilligan and Hoddinott, 2007), appropriate forms of cash versus in-kind transfers (Basu, 1996), and the efficacy of conditionality. However, much of the evidence fails to take endogeneity of programme placement and participation into account, and the empirical findings are far from unequivocal.

Sulaiman (2010) estimates the welfare effects of food aid in the post-conflict setting of Southern Sudan using an RCT approach. Food aid resulted in a significant negative impact (13%) on per capita household income, but no effect on working hours or economic activities by adult members was registered. The decline in income is thus mostly due to a reduction in child labour. There is also a positive effect on school attendance by girls (about 10 percentage points) and an improvement in their housing status. No indications are found of crowding out private transfers to participants, but there is a small but significant impact on the private transfers given out by the participants.

Evaluation studies at project level usually find considerable substitution effects occasioned by targeted interventions. Ruben and Fort (2012) analyse the impact of certification on the welfare of coffee farmers in Peru, and find no significant increase in net household income. Although coffee yields and prices substantially improved, income derived from other household activities decreased since farmers shifted land and labour away from food crops and off-farm work. This illustrated that targeted aid programmes provide incentives for concentrating efforts and resources towards the supported activity but might lead to other activities being neglected. Interestingly, the expenditure effects of the programme were still positive: certified farmers devoted a lower share of their income to direct consumption and healthcare and were able to invest more in housing and education. The enhanced income security might thus provide incentives for shifting expenditures from consumption towards investment.

Similar conclusions were drawn from a recent evaluation of the Millennium Villages in Kenya (Wanjala and Muradian, 2011). Using household survey data from the Millennium Village, Sauri, and a propensity score matching methodology, this paper analyses the impact of the Millennium Village Project (MVP) interventions on agricultural productivity and income. The results show a significant increase in agricultural productivity but an insignificant income effect, which can be attributed to small land sizes and over-reliance on agriculture. The results indicate the need to diversify economic activities and point to a revision of the simple assumptions regarding the relationship between productivity and income.

These and other micro studies show that there is a need for a thorough evaluation of the effects and impact of aid programmes on different components and indicators of client performance. Substitution effects are easily overlooked if single result indicators are used. More attention is therefore required to gain an understanding of the behavioural reactions exhibited by the aid receiver and the likelihood of crowding-in and crowding-out responses.

5.5. Creating spillover effects

International aid is often conceptualised as an "input" capable of generating specific outcomes and results, but little consideration is given to spillover effects on other sectors or activities and/or external agents. When focussing on one specific area of intervention, insights into both positive and negative spillovers might easily be lost. Yet, if aid is considered as a "catalyst", spillover effects are indeed a major outcome.

Spillover effects are the externalities of activities or processes that influence parties who are not directly involved in these activities or processes. Different types of spillovers can be distinguished: geographical or spatial spillovers (caused by geographical proximity), technological spillovers (diffusing experience, skills and knowledge), institutional spillovers (on the functioning of local organisations) and behavioural spillovers. Many spillovers occur with inter-temporal differences and may materialise only after considerable time lags. The empirical literature suggests that the magnitude of the spillovers depends on the nature of input-output linkages, the technological complexity and type of commodities that are sourced, and the relationships within chains or networks of exchange.

Much of the research on spillover effects has been carried out within the framework of private enterprises that make (foreign) direct investments with the potential to promote broader economic growth (Blalock and Gertler, 2008). Panel data are however required in order to shed light on the likelihood of reverse causation. Moreover, spillovers may also be caused by differences in the institutional and legal environment. Madariaga and Poncer (2007) rely on sub-national level data across cities to estimate a dynamic panel growth equation that takes into account the issues of spatial dependence in China. Their analysis makes it possible to determine whether FDI is characterised by a substitution or complementary pattern across Chinese cities. Results show that economic growth responds positively to capital inflows received locally as well as in proximate locations. A 50% increase in real per capita income in surrounding cities results in a 10% increase in local income. In a similar vein, Spencer (2008) focuses on knowledge spillover from foreign FDI to identify positive horizontal spillover to indigenous firms. These spillovers include demonstration effects, local linkages, employment effects and competition effects that result in higher resource productivity. It is suggested that long-run spillovers are usually larger than short-run effects, while the magnitude of spillover effects tends to be higher if foreign management is willing to engage in local strategic alliances. Another study by Jordaan (2011) in Mexico confirms that FDI firms generate substantially larger local dynamic impact through backward linkages: foreign-owned firms apply more pressure on their suppliers to improve and are also considerably more involved in providing various types of technological and

organisational support. Havranek and Irsova (2011) recently published a meta-review on vertical spillovers from FDI and conclude that average spillover to suppliers is positive and economically significant, whereas the spillover to buyers is negligible. Greater spillovers are experienced by countries that have underdeveloped financial systems and are open to international trade. Greater spillovers are generated by investors from distant countries, and who have only a slight technological edge over local firms.

The static and dynamic impact of aid on growth and development is frequently analysed using a similar framework. Some useful conclusions can be derived by comparing potential spillovers from different types of aid. Many empirical studies have shown that infrastructure contributes to economic growth and poverty reduction in developing countries, even though the sustainability of such investments largely depends on institutional spillover effects. Technical cooperation grants may create substantial knowledge spillovers that make a key contribution to factor productivity growth (Sawada *et al*., 2010). However, these learning-by-doing effects are usually weaker than the effects procuded through more external openness and free trade.

Regional pubic goods – such as regional banks, feeder roads, waterways, power networks, natural resource management and local security systems – are strongly complementary to private investment and create large spillovers. The same holds true for international public goods – climate, cross-border disease control, financial stability and security – that are not subject to pricing and therefore suffer from structural undersupply. International cooperation is nowadays frequently suggested as means of guaranteeing adequate provision of these public goods and thus of safeguarding the positive spillovers.

Many impact evaluations of development programmes do not usually explicitly take into account externalities for non-participants. RCTs that use the random characteristics or eligibility criteria of the programme evaluated are more able to identify the spillover effects. However, if implementation is not random (e.g. targeting) or participation is voluntary and open-ended, identifiying the treatment effect becomes somewhat tricky. In this case, a conventional comparison of participants with the control group may not be able to measure the spillover effects within the control village. This could lead to an underestimation of the programme's effects if the outcomes of participating households are compared with the improved outcomes of non-participants. However, once provision for such "leverage" or "catalytic" effects is incorporated into the evaluation design, results reveal that significant spillovers occur.

Janssens (2005) not only finds direct effects on the immunisation rates for participants' children but also significant spillover effects on the immunisation rates for non-participants' children in rural India. The impact of interventions might be substantially underestimated if such external effects are not taken into account. In the case of immunisation, the programme externalities for non-participants are 40-50% of the direct programme effect on participants. Likewise, the programme spillover effect on the preschool enrolment

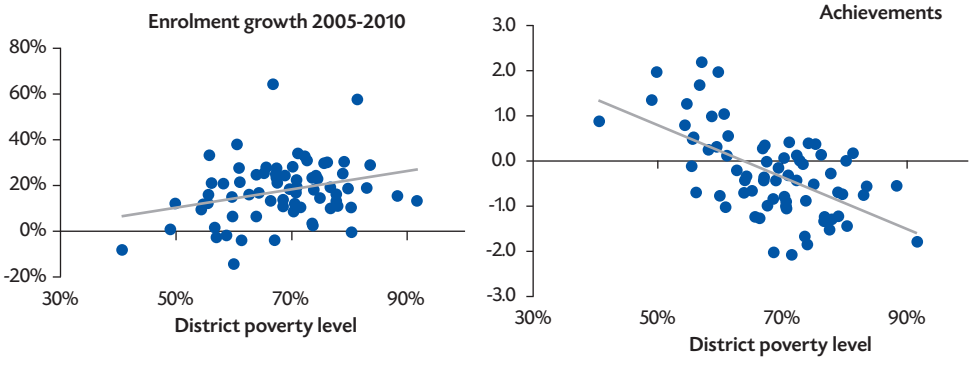
of non-participants is equal to 54% of the magnitude of the direct programme effects. Finally, programme spillovers on school enrolment of non-participants represent 49% of the total impact on participants (but not significant for the sub-sample of boys). In a similar vein, Kremer *et al.* (2009) find that merit grants for education to adolescent girls in Kenya also have large positive spillovers on non-clients (boys) and even on their parents.

Spillovers are also frequently acknowledged as a key component in strategies to address development in a fragile state context. Since the external costs of fragility to neighbouring countries are extremely high, there are sound reasons to focus development assistance on "turning around" failed states in pre- and post-conflict periods.[44] Both technical assistance and other types of aid show significant effects on the time it takes for an incipient turnaround to become a sustained turnaround. For both, however, the relationship is non-linear. Technical assistance is subject to diminishing returns with an optimal amount around 5% of GDP. The non-linearity for financial aid indicated that small amounts may actually slow down the turnaround process, while only very big aid volumes (>30% of GDP) prove effective (Chauvet and Collier, 2004). Such thresholds imply that due attention should be given to donor alliances and aid harmonisation.

Similar externalities could also be acknowledged in analyses regarding the effectiveness of macroeconomic (budget) support. Instruments for general and sectoral budget support are explicitly designed to generate spillovers within and between sectors and to create leverage for more cost-efficient service provision. Donor engagement in resource and risk pooling has

Figure 6

Impact of budget support on basic education (2005-10)

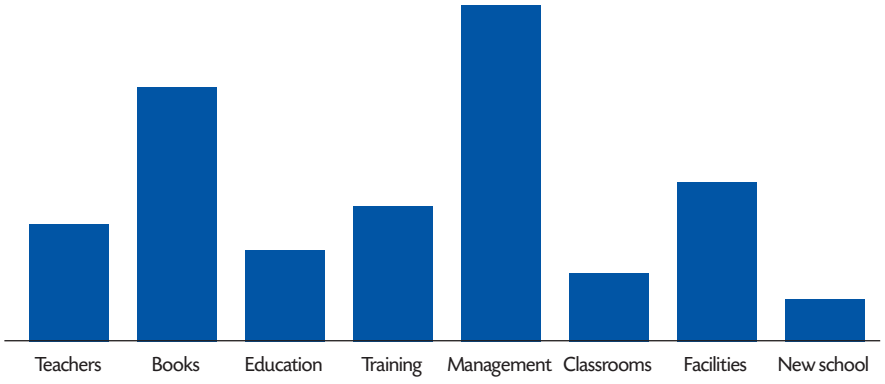


Source: IOB (2011).

[44] Chauvet and Collier (2004) estimate that the average cost in terms of net present value of just one single country falling into conflict status is USD80 billion, (i.e. larger than the world's annual aid budget to all countries). The typical neighbour loses 1.6 percentage points of its growth rate if its neighbour is a failing state.

Figure 7

Cost-effectiveness of different instruments



Source: IOB (2008).

contributed to increased discretionary expenditure and enhanced allocative efficiency in national budgets. Such results are sometimes difficult to discern, given the usual trade-off between increased access to public services and improved outcomes of the system. The IOB evaluation of budget support to basic education in Zambia shows that the increase in enrolment in poorer districts absorbs most of the improvements in learning achievements (see Figure 6).

For an adequate understanding of the potential dynamics of spillover effects from aid, it is necessary to gain deeper insights into the micro-macro linkages. Building schools, improving classrooms and even teacher training nowadays yield limited net effects. Learning achievements increase most with investments in school management, which guarantees better resource coordination (the interaction between classrooms, availability of books and teacher attendance) and thus controls premature drop-out (see Figure 7). Consequently, payment for delivery-based educational finance systems might be considered as cost-efficient procedures to create important institutional and behavioural spillovers, but only if combined with adequate monitoring and inspection procedures.

5.6. Donor roles under the new aid architecture

The international aid architecture is subject to sweeping change due to the fact that current development aid only represents a minor resource flow to developing countries (compared to FDI and remittances), and also because of the recognition that development efforts are only partly financed by aid and require substantial national contributions (from both governments and citizens).

The proliferation of different aid modalities and the fragmentation of donor support implies that due attention should be given to the choice and selection of the most appropriate aid transfer channels (bilateral, multilateral, private or civic aid) and to the preferred type of delivery mode (cash, in-kind, loans, conditional transfers, etc.) in order to foster sustainable growth and effective poverty alleviation. We have therefore discussed here some key mechanisms that determine which type of aid works. It is likely that the potential positive and adverse effects of aid are mainly the consequence of *how* aid is given.

Much traditional aid is provided under the exogenous growth paradigm, focussing on local savings or investment gaps, as inspired by Harrod-Domar growth models. Attention to resource complementarities and substitution effects increased with the introduction of Solow-type models, which account for decreasing returns and take into consideration the effects of technological change and the different stages of growth. Spillover effects have become more important in endogenous growth models, as developed by Sala-i-Martin and Barro among others. These models focus on the role of local institutional conditions and human capital in technological change processes.

Any overall appraisal of the specific role and potential of different aid modalities that take advantage of the dynamic implications of complementarities, substitution options and spillover effects is clearly beyond the scope of this paper. It seems possible, however, to outline some of the likely effects of three innovative aid modalities: (1) public-private partnerships (PPPs) that are set up to involve the private sector more directly in development programs; (2) civil society support funds (CSFs) that provide donor support for civic engagement with local or national development efforts; and (3) multi-donor trust funds (MDTF) created to mobilise support from different donors towards specific development issues. We briefly discuss the pros and cons of these aid modalities against the background of our earlier appraisal of the factors influencing development effectiveness (see Table 3). Public-private partnerships are usually highly appreciated for their capacity to generate resource complementarities and subject to very few direct crowding-out effects. However, it is extremely difficult to prove the (*ex-post*) additionality of PPPs, and

Table 3 *Channel comparison*

	Complementarities	Substitution Effects	Spillovers
Pubic-private Partnerships	High	Medium	Medium/ Low
Civil Society Support Funds	Medium	Low/ Medium	Medium High
Multi-Donor Trust Funds	Low	Medium/ High	Medium

substitution effects are likely to be relevant. Moreover, there is still scarce evidence of positive spillover effects produced by PPPs (and there are some indications of negative spillover in terms of critical employment conditions). Civil society funds face low substitution risks in remote areas and tend to be more complementary in sectoral programmes that involve several donors. Spillovers of CSF programs depend on location and entrance costs. Multi-donor trust funds are usually highly problem-focussed but characterised by limited complementarities. Crowding-out from general public budget support has been frequently mentioned. Spillover effects strongly depend on linkages with local governance structures.

The emergence of these new aid modalities makes it all the more urgent to shift attention beyond the direct registered effects when analysing impacts. This may have profound implications for the way impact studies are designed and conducted, both with respect to sample selection and the appropriate indicator framework. In addition, distributional effects (equity) and behavioural change become major components of the analysis. The evaluation frontier is therefore moving towards an analysis of the interaction effects between aid flows and complementarities with local resources, as well as the response of local agents to this set of incentives.

The appraisal of new aid modalities not only depends on their potential impact, but is also related to aspects of aid management, delivery and administration. The aid modalities discussed earlier require different types of donor support and their transaction costs are likely to vary. In a similar vein, due attention should be given to the prospects for sustainability (e.g. the likelihood that the programme will continue after closure of external funding). Moreover, prospects for sequencing aid and tailoring support to local opportunities and conditions depend on the degree of ownership and the alignment with priorities and programmes undertaken by recipient entities.

References

Abdulai, A., C.B. Barrett and J. Hoddinott (2005), "Does food aid really have disincentive effect? New evidence from Sub-Saharan Africa", *World Development*, 33(10): 1689–1704.

Albarran, P. and O.P. Attanasio (2002), "Do Public Transfers Crowd Out Private Transfers? Evidence from a Randomized Experiment in Mexico", WIDER Discussion Paper No. 2002/6, Helsinki.

Arndt, C., S. Jones and F. Tarp (2010), "Aid and Growth - Have We Come Full Circle?", *Journal of Globalization and Development* 1(2), Article 5. DOI: 10.2202/1948-1837.1121.

Barakat, S. (2009), "The failed promise of multi-donor trust funds: aid financing as an impediment to effective state-building in post-conflict contexts", *Policy Studies* 30(2): 107–126.

Barker, O. (2011), *Can aid work?*, Center for Global Development, Washington DC.

Basu, K. (1996), "Relief programs: When it may be better to give food instead of cash", *World Development*, 24(1): 91–96.

Blalock, G. and P.J. Gertler (2008), "Welfare gains from Foreign Direct Investment through technology transfer to local suppliers," *Journal of International Economics*, 74(2): 402–421.

Bourguignon, F. and M. Sundberg (2007), Aid Effectiveness–Opening the Black Box, *American Economic Review*, 97(2): 316–321.

Boyer, R. (2005), "Coherence, Diversity, and the Evolution of Capitalism; The Institutional Complementarity Hypothesis", *Evolutionary and Institutional Economics Review*, 2(1): 43–80.

Center for Global Prosperity (2010), Index of Global Philanthropy and Remittances, Hudson Institute, Washington D.C.

Chauvet, L. and P. Collier (2004), *Development Effectiveness in Fragile States: Spillovers and Turnarounds*, Centre for the Study of African Economies, Oxford.

Cordella, T. and G. Dell'Ariccia (2003), "Budget Support versus Project Aid", IMF Working Papers 03/88, International Monetary Fund, Washington D.C.

Das, J., S. Dercon, J. Habyarimana and P. Krishnan (2004), "Public and Private Funding of Basic Education in Zambia: Implications of Budgetary Allocations for Service Delivery", Africa Region Human Development Working Paper 62, World Bank, Washington D.C.

de Kemp, A., J. Faust and S. Leiderer (2011), "Between high expectations and reality: an evaluation of budget support in Zambia, Synthesis Report", BMZ / Ministry of Foreign Affairs / Sida, Bonn / The Hague / Stockholm.

Dercon, S. and P. Krishnan (2003), "Risk sharing and public transfers", *Economic Journal*, 113(486): C86–C94.

European Commission (EC) (2010), *Europeans, development aid and the Millennium Development Goals*, Special Eurobarometer No.352, TNS Opinion & Social, Brussels, Belgium.

Feyzioglu, T., V. Swaroop and M. Zhu (1998), "A panel data analysis of the fungibility of foreign aid", *World Bank Economic Review*, 12: 29–58.

Fort, R. (2008), "Assessing the impact of rural land titling in Peru: The case of the PETT program", paper to be presented at the World Bank Conference on New Challenges for Land Policy and Administration, 14-15 February 2008, Washington D.C.

Foster, M. and J. Leavy (2001), "The Choice of Financial Aid Instruments", ODI Working Paper 158, Overseas Development Institute, London.

Frot, E. and J. Santiso (2008), *OECD Development Aid and Portfolio Funds: Trends, volatility and Fragmentation*, Working Paper No. 275, OECD, Paris.

Fruttero, A. and V. Gauri (2005), "The Strategic Choices of NGOs: Location Decisions in Rural Bangladesh", *The Journal of Development Studies* 41(5): 759–787.

Gilligan, D.O. and J. Hoddinott (2007), "Is there persistence in impact of emergency food aid? Evidence on consumption, food security, and assets in rural Ethiopia", *American Journal of Agricultural Economics*, 89(2): 225–242.

Havranek, T. and Z. Irsova (2011), "Estimating Vertical Spillovers from FDI: Why Results Vary and What the True Effect Is", Working Paper 2011/07, Czech National Bank, Research Department.

Hudson, D. and J. van Heerde (2010), "A mile wide and an inch deep: Surveys on Public Attitudes towards Development Aid", available at:
http://davidhudson.files.wordpress.com/2009/03/ hudson-van-heerde-mile-wide-inch-deep-17-feb-2010.pdf.

IOB (2008), "Impact evaluation Primary Education in Uganda", IOB Evaluation No.311, Ministry of Foreign Affairs of the Netherlands, The Hague.

IOB (2011*a*), "Unfinished business: Making a difference in basic education. An evaluation of the impact of education policies in Zambia and the role of budget support", Ministry of Foreign Affairs of the Netherlands, The Hague.

IOB (2011*b*), "Education matters: Policy review of the Dutch contribution to basic education 1999–2009", IOB Evaluation No.353, Ministry of Foreign Affairs of the Netherlands, The Hague.

IOB-BMZ (2011), "The risk of vanishing effects - Impact evaluation of drinking water supply and sanitation programmes in rural Benin", MinBuza/IOB, Ministry of Foreign Affairs of the Netherlands, The Hague.

Janssens, W. (2005), "Measuring Externalities in Program Evaluation", Tinbergen Institute Discussion Paper TI 2005-017/2, Amsterdam.

Jordaan, J.A. (2011), "FDI, local sourcing, and supportive linkages with domestic suppliers: The case of Monterrey, Mexico", *World Development*, 39(4), 620–632.

Ketkar, S. and D. Ratha (2009), *Innovative Financing for Development*, World Bank, Washington DC.

Kindornay. S. (2011), "From Aid to Development Effectiveness", working paper, The North-South Institute, Ontario.

Knack, S. (2001), «Aid Dependence and the Quality of Governance: Cross-Country Empirical Tests», *Southern Economic Journal*, 68 (Oct.): 310–29.

Koch, D-J. (2009), *Aid from International NGOs: Blind Spots on the Aid Allocation Map*, Routledge, New York.

Kinsbergen, S., J. Tolsma and S. Ruiter (2011), "Bringing the Beneficiary Closer: Explanations for Volunteering Time in Dutch Private Development Initiatives", *Nonprofit and Voluntary Sector Quarterly* 40(6): doi: 10.1177/0899764011431610.

Kremer, M., E. Miguel and R. Thornton (2008), "Incentives to Learn", *Review of Economics and Statistics*, 91(3): 437–45.

Lessmann, C. and G. Markwardt (2010), "Decentralization and Foreign Aid Effectiveness: Do Aid Modality and Federal Design Matter in Poverty Alleviation?", Working Paper No. 3035, CESifo, Dresden.

Lu, C., M.T. Schneider, P. Gubbins, K. Leach-Kemon, D. Jamison & C.J.L. Murray (2010), "Public financing of health in developing countries: a cross-national systematic analysis", *The Lancet*, 375(9723): 1375–1387.

Madariaga, N. and S. Poncet (2007), "FID in Chinese Cities: Spillovers and Impact on Growth", *The World Economy*, 30: 837–862.

Masud, N and B. Yontcheva (2005), "Does Foreign Aid reduce Poverty? Empirical Evidence from Nongovernmental and Bilateral Aid", IMF Working Paper WP/05/100, International Monetary Fund, Washington D.C.

Mavrotas, G. (2005), "Aid Heterogeneity: Looking at Aid Effectiveness from a different angle", *Journal of International Development*, 17: 1019–36.

Milgrom P. and J. Roberts (1995), "Complementarities of fit: strategy, structure, and organizational change", *Journal of Accounting and Economics*, 19: 179–208.

Morrissey, O. (1990), "The impact of multilateral and tied bilateral aid on the UK economy", *Journal of International Development*, 2(1): 60–76.

Nancy, G. and B. Yontcheva (2006), "Does NGO Aid Go to the Poor? Empirical Evidence from Europe", IMF Working Paper WP/06/39, Washington DC.

Nunnenkamp, P. and H. Öhler (2009), "Aid Allocation through Official and Private Channels: Need, Merit and Self-interest as Motives of German Donors", IFW Working Paper No.1536, Kiel Institute for the World Economy, Kiel.

Ohno, I. and Y. Niiya (2004), Good Donorship and the Choice of Aid Modalities–Matching Aid with Country Needs and Ownership, GRIPS, Tokyo.

Raschky, P.A. and M. Schwindt (2008), "Aid, Catastrophes and the Samaritan's Dilemma", Working Paper No. 2008-04-15, Wharton School – University of Pennsylvania, Philadelphia PA.

Raschky, P.A. and M. Schwindt (2011), "On the Channel and Type of Aid: the case of international disaster assistance", Monash University, Caulfield.

Ricker-Gilbert, J., T.S. Jayne and E. Chirwa (2011), «Subsidies and Crowding Out: A Double-Hurdle Model of Fertilizer Demand in Malawi», *American Journal of Agricultural Economics* (in press) doi: 10.1093/ajae/aaq122.

Ruben, R. and R. Fort (2012), "The Impact of Fair Trade Certification for Coffee Farmers in Peru", *World Development*, 40(3): 570-582.

Sarkar, M.B., R. Echambadi, S.T. Cavusgil and P.S. Aulakh (2001), "The influence of complementarity, compatibility, and relationship capital on alliance performance", *Journal of the Academy of Marketing Science,* 29(4): 358–373.

Sawada, Y., A. Matsuda and H. Kimura (2010), "On the role of technical cooperation in international technology transfers", *Journal of International Development*, DOI: 10.1002/jid.1685.

Schiffman, J. (2006), «HIV/AIDS and the rest of the global health agenda», (editorial) *Bulletin of the World Health Organization*, 84(12): 923.

Schulpen, L., B. Loman and S. Kinsbergen (2011), "Worse than Expected? A comparative analysis of donor proliferation and aid fragmentation", *Public Administration and Development*, 31(5): 321–339.

Shields, M.P. (2007), "Foreign aid and domestic savings: the crowding out effect", Discussion Paper No. 35/07, Department of Economics, Monash University.

Spencer, J.W. (2008), "The Impact of Multinational Enterprise Strategy on Indigenous Enterprises: Horizontal Spillovers and Crowding Out in Developing Countries", *Academy of Management Review*, 33(2): 341-361.

Sulaiman, M. (2010), "Incentive and crowding out effects of food assistance: Evidence from randomized evaluation of food-for-training project in Southern Sudan", EOPP/2010/19, London School of Economics and BRAC.

Feyzioglu, T., V. Swaroop and M. Zhu (1998), "A Panel Data Analysis of the Fungibility of Foreign Aid", *World Bank Economic Review*, 12(1): 29–58.

Tadesse, G. and G. Shively (2009), "Food aid, food prices and producer disincentive in Ethiopia", *American Journal of Agricultural Economics*, 91(4): 942–955.

Tooley, J., P. Dixon and J. Stanfield (2008), "The impact of free education in Kenya: a case study in private schools in Kibera", *Educational Management, Administration and Leadership*, 36(4): 449–469.

Vandycke, N. (2000), "Improving Capabilities: Education"; in *Making Transition Work for Everyone: Poverty and Inequality in Europe and Central Asia,* World Bank, Washington D.C.

van Heerde, J. and D. Hudson (2010), "'The Righteous Considereth the Cause of the Poor?' Public Attitudes towards Poverty in Developing Countries", *Political Studies* 58(3): 389–409.

Vos, R., A. Bedi, P. Kimalu and D. Manda, (2004), "Achieving universal primary education: Can Kenya afford it?", Department of Economics Working Paper Series, University of Connecticut.

Wanjala, B.M. and R. Muradian (2011), "Can Big Push Interventions Take Small-scale Farmers out of Poverty? Insights from the Sauri Millennium Village in Kenya", CIDIN Working Paper 2011-1, CIDIN, Nijmegen.

White, H. (1993), "Aid and government: A dynamic model of aid, income and fiscal behaviour", *Journal of International Development* 5(3): 305–312.

6. Applying Evaluation to Development and Aid: Can Evaluation Bridge the Micro-macro Gaps in Aid Effectiveness?

Leonce Ndikumana, Department of Economics and Political Economy Research Institute (PERI), University of Massachusetts[45]

Abstract

Donors and governments in aid recipient countries are under pressure to demonstrate the effectiveness of aid, especially given the growing stress on fiscal balances in the context of the global financial and economic crisis. The evidence on aid effectiveness remains mixed at best: while individual targeted aid interventions appear to produce positive results, the impact of aid at the macroeconomic level remains limited. Furthermore, the reporting on concrete outcomes of aid interventions remains inadequate, thus perpetuating doubts as to aid effectiveness. This paper discusses these micro-macro gaps in aid effectiveness and the reporting problem. It proposes some ways in which well-designed and carefully implemented evaluations can help bridge these gaps, and how better reporting and transparency on aid results can advance the aid effectiveness agenda.

[45] The author is grateful for excellent assistance from Theresa Owusu-Danso, doctorate candidate at the University of Massachusetts

6.1. Introduction

There is growing pressure on donors and recipient governments to demonstrate the effectiveness of aid. In donor countries, taxpayers demand tangible proof of the use of the tax money channelled through national aid agencies and multilateral institutions. This pressure has been exacerbated by the adverse impact of the global financial and economic crisis on donors' fiscal balances. At the same time, populations in recipient countries are more and more openly demanding tangible development outcomes, more transparency in the management of aid and better access to reports containing systematic and objective assessments of aid effectiveness. In developing countries, along with increasing democratisation and a free press, governments are facing a growing pool of educated but disfranchised youth demanding genuine improvements in living standards.

These growing pressures and demands for transparency and aid effectiveness are further fuelled by criticisms ranging from analysts arguing that aid has no robust impact on development, to activists who are openly opposed to aid on various grounds. Furthermore, the complexity of the development process makes it difficult to track the impact of aid, which is influenced by many factors emanating from the donor side, the recipient's context and exogenous factors. In addition, the aid industry is a congested market where multiple actors pursue similar goals on the same terrain, making it even more difficult to sort out the incremental impact of aid interventions.

Nonetheless, evidence shows that development aid has produced substantial positive results at the micro level, whether at project or programme level. Well managed programmes have yielded improvements in school enrolment, access to healthcare, reforms of tax systems and other valuable outcomes. However, at the aggregate level the record remains very mixed, fuelling the debate about overall weak aid effectiveness. Bridging the micro-macro gap remains a critical challenge for the development aid community and national policy makers.

These practical challenges, criticisms, gaps between micro and macro outcomes and domestic political pressures on donors and recipient governments for more transparency on aid call for more effective mechanisms to analyse, monitor, evaluate and disseminate the concrete impacts of aid on development; in other words, there is a call for better aid effectiveness evaluation. While there has been substantial progress in evaluation methods and practice, important gaps remain and there is room for improvement. Moreover, the dissemination of impact evaluation results remains inadequate, which further perpetuates doubts about aid effectiveness. This paper argues that well designed and implemented evaluations, together with better dissemination of evaluation results, can help shed light on these gaps in aid effectiveness. The paper thus emphasises two problems:

the dichotomy between micro-level and macro-level aid effectiveness; and the lack of transparency and inadequate reporting on the concrete impacts of aid. These problems are at the heart of the concerns regarding aid effectiveness in the aid community as well as in recipient countries.

Following this introduction, the paper provides a brief review of the mixed record of aid effectiveness in Section 2. Section 3 highlights the problems at the origin of the micro-macro dichotomy in aid effectiveness. Section 4 discusses the role that evaluation can play in bridging these gaps, and Section 5 concludes.

6.2. Aid effectiveness: a less than stellar record

6.2.1. A backdrop of rising aid volumes

The debate on aid effectiveness and evaluation is taking place in the context of an upswing of aid flows to developing countries. Following a steady decline in the 1990s, total aid by the Development Assistance Community (DAC) member countries has increased substantially since the turn of the current century. Between 1990 and 2009, total aid to all developing countries by DAC donors rose from USD119.9 billion to 165.3 billion, a 37.8% increase. Based on the trough levels of 2000 (USD84.7 billion), this represents a doubling of the volumes of aid over a decade (Table 4).

Table 4 *Real annual flows of official aid (constant 2009 US dollars, billions)*

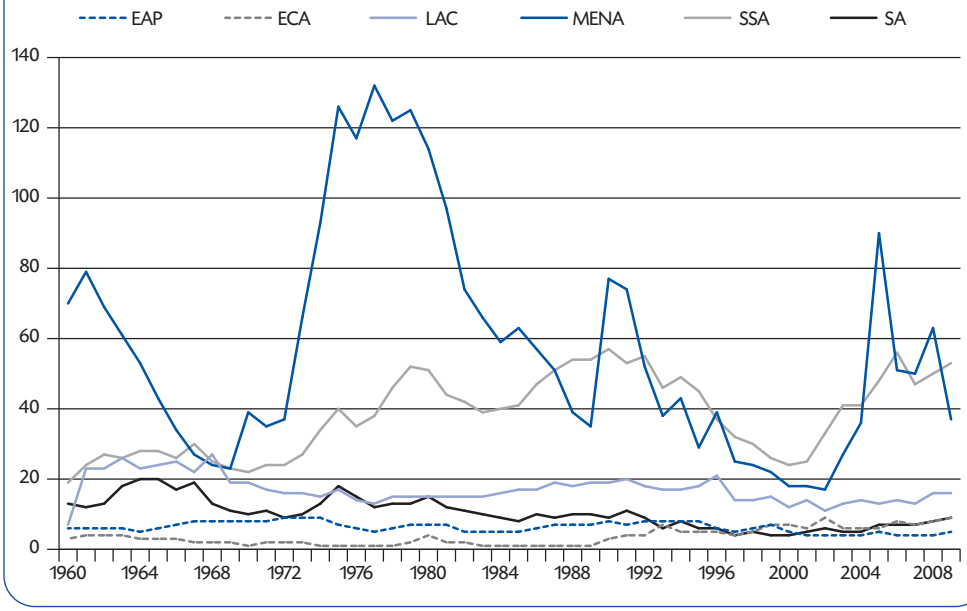
Year	Total	Africa	Latin America	Asia	Europe	Oceania	Unspecified
1960	30.8	9.5	1.6	15.7	2.8	0.2	1.0
1970	44.9	9.3	5.7	18.5	1.0	1.5	1.9
1980	111.4	27.1	5.8	35.3	3.1	2.7	12.8
1990	119.9	41.2	8.6	29.5	2.3	2.3	9.6
2000	84.8	19.2	6.0	19.7	4.6	1.0	11.1
2009	165.3	47.7	9.1	38.6	5.8	1.6	24.9
Change 1990-2009 (%)	37.8	15.8	5.9	30.7	147.6	-26.9	159.1
Change 2000-2009 (%)	95.0	148.7	50.8	95.7	25.0	62.0	122.9

Source: DAC database (online). Nominal values are deflated into real values using the US CPI index.

In per capita terms, the upward trend is most notable in sub-Saharan Africa, Asia and Latin America (Figure 8). The substantial increase in aid to these regions since 2000 has been credited for contributing to high growth in the pre-crisis period.[46] Aid per capita in sub-Saharan Africa more than doubled between 2000 and 2009, rising from USD24 to USD54. However, the 2009 levels are still below the peak of USD57 per capita reached in 1990.

Figure 8

Real aid per capita by region, 1960-2009 (constant 2009 US dollars)



Source: DAC database (online). Nominal values are deflated into real values using the US CPI index. SSA = sub-Saharan Africa; MENA = Middle-East and North Africa; SA = South Asia; LAC = Latin America and the Caribbean; EAP = East Asia and Pacific; ECA = Europe and Central Asia.

Despite the considerable increase in the volumes of aid over the recent years, the quantity of aid remains inadequate relative to the financing needs of developing countries as well as relative to donors' targets. The High-Level Plenary Meeting on the Millennium Development Goals (MDGs) held in New York in 2010 under the theme of "keeping the promise" soberly lamented the fact that donors had not kept their promise of increasing aid delivery (UN MDG Task Force, 2011). The report of the MDG Task

[46] Various reports by the multilateral development institutions have listed increasing volumes of aid as one of the key drivers of the high growth in Africa during the years leading to 2008–09. These include the *African Economic Outlook* (by the African Development Bank, OECD, UNECA and UNDP), UNECA's *Economic Report on Africa*, UNDESA's *World Economic and Social Prospects*, and the IMF's *World Economic Outlook*.

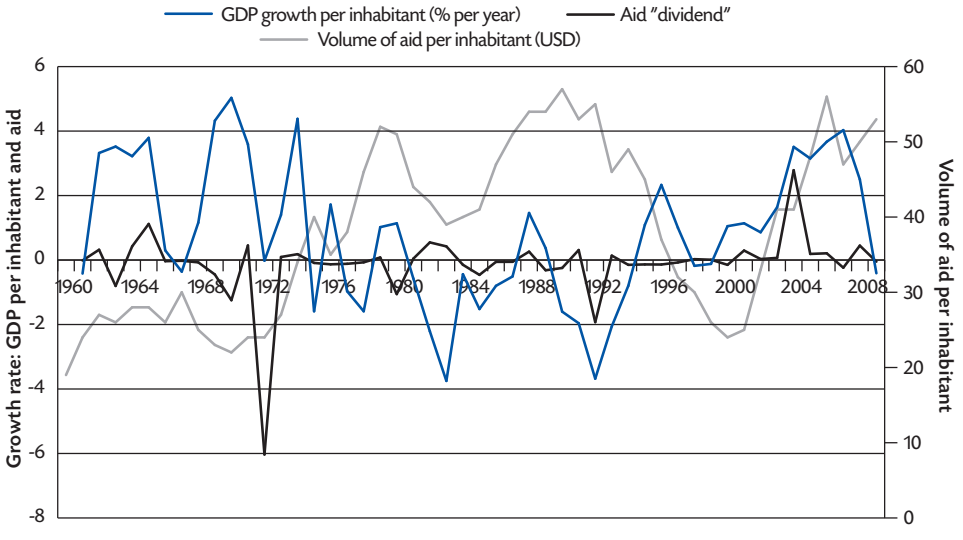
Force found that, while official aid had reached a record high of USD129 billion in 2010, this represented only 0.32% of the gross national income (GNI) of DAC members. Only five countries have met the UN target of 0.7% of GNI in official aid.[47] The report noted a large gap of USD153 billion in 2010 in actual aid delivery. Aid delivery to Africa in 2010 was USD15 billion (in 2004 dollars) below the pledges made in 2005 at Gleneagles (UN MDG Task Force, 2011: 15).

At the same time, developing countries are facing large financing gaps in economic infrastructure and social sectors. It is estimated that Africa faces an annual gap of USD48 billion in infrastructure financing alone. In 2008, the MDG Africa Steering Group Report (2008) concluded that for African countries to reach the MDGs, public external financing would need to be scaled up by about USD72 billion per year until 2010. Actual disbursements fall far below these targets.

But most importantly, despite the fact that the volumes of aid to developing countries in general have increased over the past years, the record of the impact of aid on

Figure 9

Growth gains from aid in SSA: GDP growth/aid growth



Source: DAC database (online). Nominal values are deflated into real values using the US CPI index. Aid dividend is proxied by the ratio of real GDP growth to the growth rate of real aid per capita.

[47] The five countries are: Norway (1.10% of aid/GNI), Luxemburg (1.05%), Sweden (0.97%), Denmark (0.91%), and Netherlands (0.81%). (Source: OECD-DAC online database).

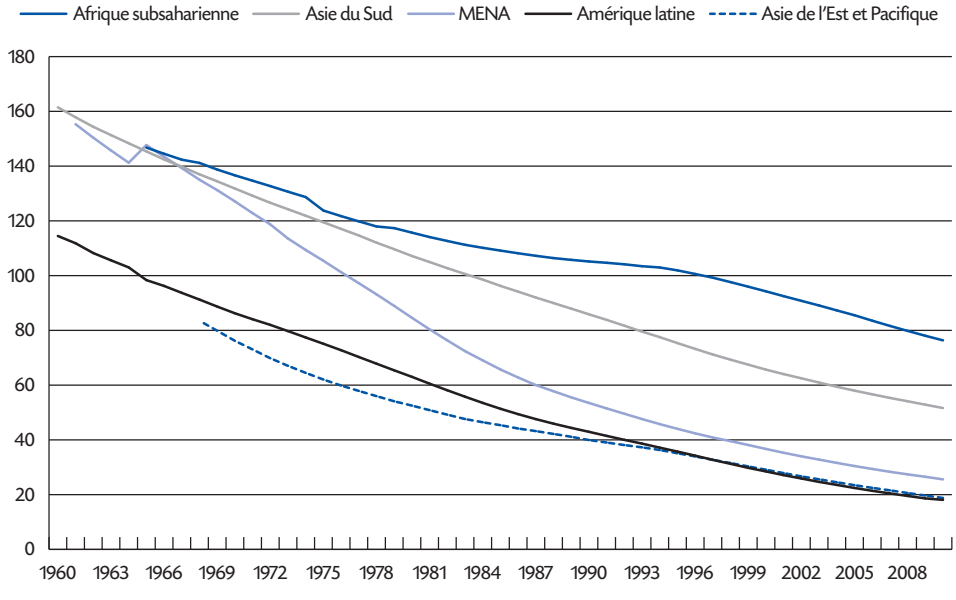
development remains rather wanting. Growth in sub-Saharan Africa remains below the levels needed to reach national development targets; growth is volatile and has generated inadequate job creation. The credit given to aid for stimulating the recent resurgence of growth in Africa is often exaggerated. Over the long run, the gains from growth are limited. As can be seen on Figure 9, growth elasticity of aid has been low and flat. The recent upswing in aid has only yielded a short-lived spike in "aid dividend", reverting to a stagnant mean.

The gains from aid in terms of social development are also less than satisfactory. While aid has supported important projects in education and health, the overall impact remains inadequate. The deficiencies are most notable in the case of Africa, the region pointed out as having received relatively higher volumes of aid. Infant mortality has declined much more slowly in Africa than in the other developing regions (Figure 10) and many African countries are not likely to reach the MDG target for this development objective.

Although the debate on aid effectiveness has heated up in recent years, efforts to assess the effectiveness of aid date from as far back as the 1960s (Roodman, 2007*a*). Doucouliagos and Paldam (2005) provide a comprehensive literature review, pointing out the "sad results" of four decades of research on the theme. Over the years, the

Figure 10

Infant mortality by region (per 1,000 live births)



Note: MENA = Middle-East and North Africa.6.2.2. Disputed evidence at the macro level but more encouraging results at the micro level

work on aid effectiveness has been reviewed several times, with conclusions ranging from extreme optimism to quasi-militant pessimism.[48] But the literature shows a clear demarcation between the findings at the macro level, where the results are mixed, and results at the micro level, where the evidence is much more encouraging regarding the gains from aid. The sections below discuss the evidence at the two levels in turn.

Mixed evidence at the macro level

The ultimate objective of aid is to contribute to improved economic performance and the wellbeing of the population through the provision of financial resources and technical assistance. Naturally, governments and the public from donor and recipient countries expect to see improvements in indicators of economic performance and wellbeing as returns to aid. This explains why the attention in the aid effectiveness analysis has been focused on national level indicators, mainly economic growth, health outcomes and human capital development. Implicitly, the analyst assumes, or rather hopes, that the impacts of aid at the micro level, where the action takes place, somehow translate into macro level impacts at the national level. But the process of aggregation of micro level outcomes into the macro level impacts remains a black box.

The literature on the macro level impact of aid falls into three camps, with a limited number of agnostics along the spectrum of aid effectiveness beliefs: aid works; aid does not work; aid works under certain conditions (but it works after all). The first camp claims that aid works, and the only concern is that there is not enough aid and that it may not reach the intended recipients. The most vocal advocates in this camp include Jeffery Sachs, who has argued forcefully for a "big-push"-led growth financed by scaling up aid to developing countries (Sachs, 2005).[49] Sachs and his colleagues argue that external development assistance can help break the poverty trap and that scaling up aid is the exit strategy from the poverty trap. The following excerpt says it all in reference to Africa:

> If Africa is caught below the threshold level of infrastructure, and therefore is stuck in chronic low or negative growth, the main policy implication is to raise capital above this threshold … We propose to increase the capital stock in one step, as it were, through a large, well-targeted infusion of foreign assistance. In other words, we are arguing not for endless flows of increased aid, and not for aid as simple charity, but rather for increased aid as an exit strategy from the poverty trap. For those who fear that aid increases dependency, our response is that aid that is ambitious enough would actually end Africa's dependency. Moreover, we see no other likely successful strategy for ending Africa's poverty trap. (Sachs *et al.* 2004: 144)

[48] Key reviews of the literature on aid effectiveness include: Mosley (1980); Hansen and Tarp (2000); Clemens *et al.* (2004); McGillvray *et al.* (2005); Roodman (2007).

[49] See Easterly (2006*a*) for a critical review of Sach's argument for a big-push approach to development assistance.

This camp includes analysts who argue that aid has been effective in stimulating growth (see Hansen and Tarp 2000, 2001). Besides the quantitative analysis supporting the aid effectiveness view, there is a large literature from the activist world (NGOs and civil society organisations) calling for a scaling-up of aid.

The second camp argues that aid works, but that it works under certain specific condition. One of the most widely cited studies in this group is by Craig Burnside and David Dollar (2000), which claimed that aid works but only in a good policy environment. This study generated considerable debate and controversies. Some analysts questioned the robustness of the results and the merit of the methodology (see Roodman, 2007*b* for a review). Further investigations refuted the results as too fragile, not robust to sample selection and subject to particular specification of the empirical model (Dalgaard and Hansen, 2001; Easterly *et al.*, 2003). For instance, the interaction of aid and policy and the particular coding of the good policy indicator are found to be the key drivers of Burnside-Dollar results. Outside of academia, the concerns with the Burnside and Dollar proposition were about its policy implications. The proposition implied that aid should go to countries with demonstrated evidence of good policy; that is, aid should be conditioned to good policies. This reopened a can of worms in the debate on aid conditionality. Most fundamentally, it meant that, given that low-income countries and especially those coming out of conflicts also have weak institutions and policy frameworks, an application of the Burnside-Dollar proposition would leave these countries as aid orphans and trapped into a low-aid–poverty vicious circle.

A number of other analysts have supported the view that aid is effective under certain conditions. Noteworthy studies include Collier and Dollar (2004), who argue that aid effectiveness requires good government institutions. Similarly, Svensson (1999) argues that aid is effective only in democracies. Collier and Dehn (2001) posit that aid can be effective in countries experiencing shocks, but point out that aid effectiveness requires good policies. Patrick Guillaumont and his colleagues argue that aid helps absorb economic and natural shocks, and strongly advocate for allocating official development aid on the basis of economic vulnerability.[50]

Within this strand of conditional aid effectiveness literature, Dalgaard *et al*. (2004) controversially suggested that aid works outside of the tropics but not within the tropics! The study is empirically fragile, the results being driven by a few countries with specific features, namely Botswana, Egypt, Jordan and Syria (Roodman, 2007*b*). This kind of conclusion feeds the usual deterministic view of development that tends to attribute underdevelopment to fixed factors such as geography. But this view is tenuous; it fails to explain, for example, why geography would prevent Burundi from developing while Switzerland developed, although both countries are landlocked and small.

[50] Guillaumont (2007, 2009, 2010); Guillaumont and Chauvret 2001; Guillaumont and Guillaumont-Jeanneney (2009); Guillaumont and Simonet (2011).

There is a smaller strand of the literature that argues that aid simply does not work, conditionally or absolutely. Rajan and Subramanian (2005) challenge the robustness of the results in studies that conclude that aid works even if it is conditional on good policies. They conclude that aid does not have any statistically consistent effect of growth and that, even in cases where it may exist, the effect is too small to be statistically observable.

William Easterly argues that aid works only if it is well targeted and aligned with individual country's cultural, social and economic conditions (Easterly, 2006*b*). He is critical of large-scale or grand-scheme types of aid interventions, or what he calls a "transformational" approach to aid. The problem is not the money; it is whether the funds are used to meet the specific needs of the intended recipients. Easterly suggests that well-managed aid produces positive results at the micro level in areas such as education and health. He thus favours the "marginal" approach with small-scale targeted interventions (Easterly, 2009).

Overall, the review of the literature suggests that the evidence on the macro-level effectiveness of aid remains mixed with no apparent movement towards any consensus. Now we turn to the micro level impact of aid, where the results are much more promising.

More encouraging evidence on aid effectiveness at the micro level

The key challenge to efforts to document and quantify the effectiveness of aid at the macro level is that macroeconomic outcomes are the result of a multiplicity of factors, many of which are unrelated to aid, and some of which can affect the effectiveness of aid either positively or negatively. Economies are complex systems where virtually everything depends on and influences everything. Disentangling the impact of a single factor such as aid on macroeconomic outcomes like growth, human capital, health, etc., is a daunting exercise both conceptually and empirically. Moreover, project level interventions may produce macro level results, but these would be observed only several years down the road (Radelet and Banana, 2004).

Moreover, and most fundamentally, aid is only an instrument used to achieve ultimate macro level goals. For the instrument to have an impact on the ultimate goal, a long chain of causalities needs to hold systematically. A chain is as strong as its weakest link; if one node in the chain of causalities fails, then the final result is compromised. For example, the ultimate goal of aid interventions in education is to increase human capital, which in turn would increase growth and generate improvements in the overall wellbeing. So donors finance school construction with the hope that the recipient country will reap future benefits in terms of improved human capital and higher growth. However in practice, for the final result to materialise, not only does aid need to be spent and used diligently, but also agents' behaviour needs to respond appropriately and significantly along the way. So, effectiveness of aid operates at multiple levels and it is the aggregation of the intermediate levels of effectiveness that determines effectiveness at the macro level. Using the example of aid to education through construction of schools, Roodman (2007*a*: 2) summarises

some of the questions that need to be addressed, which points to many ways in which aid effectiveness may be compromised: "Was a school built? Did children come? Did they learn? When they grew up, did they have fewer children of their own? Did they find more rewarding and productive work? Did economic output go up? Did poverty or inequality fall?"

For aid to education to have macro level effects, there are too many "ifs" that need to be satisfied. If more schools are built, school attendance will increase, literacy will increase, households will make more efficient decisions regarding matters relevant for their wellbeing, workers (educated) will be more productive, more output and income will be produced and the living standards will increase. Trying to demonstrate empirically each of these causal statements is a monumental task.

One possible solution to the challenge of demonstrating aid effectiveness is to be less ambitious in the quantitative assessment of aid effectiveness and look not at the macro outcomes but at the micro level outcomes; that is, look at narrower goals. Such an investigation typically reveals what Clemens *et al*. (2004) refer to as a "micro-macro paradox": despite the disappointing results at the macro level, there is evidence of successful targeted aid interventions at the micro level. At the sectoral level, aid has also been found to be effective, especially in the areas of education and health. Michaelowa and Weber (2006) find that aid contributes to increasing primary school enrolment. Dreher *et al*. (2007) find similar results. In the area of health, Mishra and Newhouse (2007) find that aid helps reduce infant mortality.

The reality, therefore, is that the aid landscape includes a mixture of successes and failures. The problem is the technology used to aggregate the impact of aid. This is at the root of the fact that the successes have not been able to outshine the failures to produce robust overall positive impacts of aid. There are many reasons for this. Key among these is that the aid enterprise has many structural deficiencies that undermine its effectiveness. These inefficiencies prevent the aggregation of positive micro level results into visible positive outcomes at the macro level.

6.3. Problems leading to the micro-macro paradox

This section discusses succinctly the key structural problems of aid effectiveness that may be at the origin of the micro-macro paradox. The focus is on problems that may be addressed through effective conceptualisation and implementation of evaluation.

6.3.1. A quantity and quality problem

To the extent that aid effectiveness means development effectiveness, then both the quantity and the quality of aid matter. Regarding the quantity of aid, if aid is to produce positive and visible results at the aggregate level, it must reach a minimum threshold. It has been pointed out in several studies and reports that the current levels of aid remain inadequate. They fall short of the investment gaps faced by developing countries and are below the OECD targets of 0.7% of donor countries' gross national income.

Various studies have documented large and even growing financing gaps faced by developing countries. In the case of Africa, for instance, it is estimated that to reach the high growth required to substantially reduce poverty, the continent would need to invest about USD93 billion per annum in infrastructure, including USD41 billion for the power sector (Africa Infrastructure Country Diagnostics, 2009). Currently only USD45 billion are covered, leaving a gap of USD48 billion, of which USD23 billion is required for the energy sector alone.

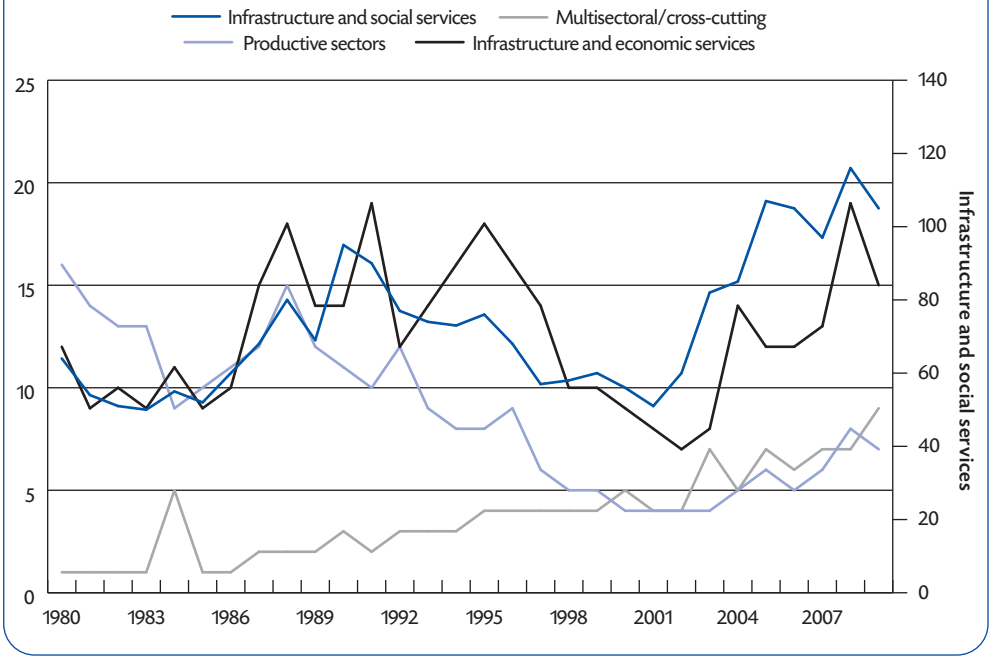
For aid to generate meaningful impacts at the macro level, the current levels would have to be substantially increased in a predictable manner to fill the investment financing gaps. Higher and more predictable funding would help achieve higher and less volatile growth, and ultimately foster social development. Improvement in aid effectiveness at the macro level is conditional on increasing the volumes of aid delivery.

The quality of aid is also essential for aid effectiveness. When aid effectiveness is defined in terms of development effectiveness rather than financial soundness or operational/process conformity, the quality of aid raises a number of issues. The two most prominent issues are allocative effectiveness and predictability. The challenge of allocative effectiveness emanates primarily from the fact that (1) resources are scarce and therefore donors have to make difficult choices on where to invest these resources; (2) there is inadequate information on effective returns to investment in various activities; and (3) there is imperfect evidence on the key drivers of growth at the country level. As a result, aid effectiveness is constrained by the fact that some resources are allocated to sectors with limited returns to investment and with little impact on growth and development.

Allocative efficiency problems are exacerbated by the lack of consistency in the decision-making of donors. Over time, donors change their aid targets and preferences, but it is not always clear whether these shifts are inspired by careful analysis of the expected relative gains from investment in various sectors. So, for example, donors have exhibited a strong bias toward social infrastructure and services, especially since 2000 (Figure 11). In contrast, aid to productive sectors declined since the beginning of the 1990s. As donors focused on poverty reduction as the ultimate goal of aid, the attention shifted to activities and sectors that were deemed closer to this objective, hence the emphasis on social sectors. This shift, however, is problematic. It has been well documented that sustained poverty reduction requires higher and sustained growth and job creation, which in turn requires adequate investments in productive sectors. Ironically, focussing on the poor by increasing spending on social services has not contributed much to reducing poverty. It is by supporting wealth and job creation through strong, sustained and broad-based growth that sustained poverty reduction can be achieved.

Figure 11

Sectoral allocation of ODA (total all donors, constant 2009 dollars, billions)



Source: DAC database (online). Nominal values are deflated into real values using the US CPI index (obtained from the IMF's International Financial Statistics).

The problem of targeting of aid is amplified by "herd behaviour" among donors, and the tendency to "follow the winner" in a context of high pressure to show results. Individual donors seek to minimise risks by avoiding untapped terrains and focussing on sectors and activities that have gained consensus among the donor community. Moreover, multilateral development institutions, which are key players in the aid landscape, do not have a genuine capacity to set their own targets. They are all accountable to the same governments of member states. As a result, the preferences of dominant donor countries permeate the strategic decisions of multilateral development institutions, so that the preferences of the latter mimic those of the former.

In addition, aid effectiveness is hampered by poor cost-effectiveness, notably due to long and cumbersome aid delivery processes. Aid for seeds and fertilisers is of little help when it is delivered after the end of the planting season. The high cost of aid delivery is also due to ineffective donor coordination in a landscape marked by a proliferation of donors and projects. This increases the burden on recipient governments called to execute, monitor, and evaluate multiple projects and dialogue with multiple donors. It is not surprising, for example, that some recipient governments occasionally call for a moratorium on donor missions during certain periods so they can get down to running their business.[51]

6.3.2. Weak additionality of aid

The limited record of aid effectiveness at macro level can also be attributed to weak additionality of official development aid. One of the reasons for the weak additionality of aid is that additionality is not integrated in the planning of aid programmes. Additionality of aid can be evaluated at three levels: financial additionality; positive spillover effects into the local economy from aid-funded projects and programmes; and technology and knowledge transfer. Financial additionality of aid stems from the role that aid plays in attracting additional public and private resources. From the public side, aid can crowd in domestic public resources by increasing the capacity to mobilise tax and non-tax revenue. A donor intervention can also crowd in external public resources by enticing other donors to co-fund programmes and projects.

From the private side, aid can play a catalytic role in attracting private financiers or by facilitating public-private partnership funding arrangements. In practice, however, instead of crowding in domestic public resources, aid often tends to have a disincentive effect on tax mobilisation (Ostrom *et al.* 2001: Xviii). Because aid is fungible with other government resources, especially in the case of budget support, high volumes of aid alleviate pressure on the government to mobilise taxes that are politically undesirable.

[51] In Tanzania for example, the government declares a mission-free month during the budget preparation process.

The ability of aid to catalyse additional public and private resources is also limited because this is not explicitly built into aid programming. When institutions make it an objective for their funds to play a catalytic role, aid indeed can crowd in substantial amounts of private finance.[52] Unfortunately, this practice is not part of the normal business of public sector aid programming among bilateral donors or multilateral development institutions.

Aid projects tend to also have suboptimal spillover effects in the local economy, which limits overall aid effectiveness. Aid programmes often remain virtual islands in the economic landscape, thus minimising their impact at the macro level.

In addition to financing, development assistance can also provide an avenue for the transfer of technological know-how from donors to aid recipients. This in turn would eventually increase productivity in recipient countries, leading to overall higher economic performance. The record of aid effectiveness in this regard is weak. The gains through technology transfer are particularly low in the case of tied aid. Despite calls to move away from tied aid, it still represents a substantial fraction of total aid for many donors, whether *de jure* or *de facto*. This further undermines aid effectiveness.

6.3.3. Failure to influence policy and institutions

There is broad consensus that institutions and good policies are important ingredients for sustained high long-run growth. Yet, the aid community has not made up its mind whether aid should be used to induce improvements in institutions and policies. For a long time, donors have had it backwards: they have conditioned aid to good institutions and policies. Given that the majority of low-income countries have weak institutions and policies, they then end up receiving less aid. As a result, poor countries are trapped in a stable equilibrium of bad institutions and low growth. Indeed, Birdsall (2007) argues that what is holding African economies in a low-growth high-poverty trap is an "institutional trap". While growth is believed to be a function of institutions, donors have primarily focused on the direct link between aid and growth, and less on institutions. One of the causes of limited performance of aid in stimulating growth is that little aid has been invested in institution building and that aid has not been sufficiently leveraged to improve the institutional framework in low-income countries.

There are reasons for this limited emphasis on using aid to develop institutions. Some of these constraints are political, whereby donors put their national strategic interests ahead of recipient countries' economic development goals (Killick, 1998; Kanbur, 2000; Mold, 2009). Thus, bad governance in recipient countries goes unchallenged, and even worse is rewarded by additional aid inflows in the name of national strategic interests. On the recipient side, there is resistance against interventions with an

[52] IFC's evaluation reports show strong additionality of private sector lending, with the most common form of additionality arising from funds mobilisation. See for example, Independent Evaluation Note No.1, 2008. "IFC's Experience and Additionality in Middle Income Countries Results and Challenges"; also see IFC (2010) "IFC's Additionality: A Primer", September 30, 2010.

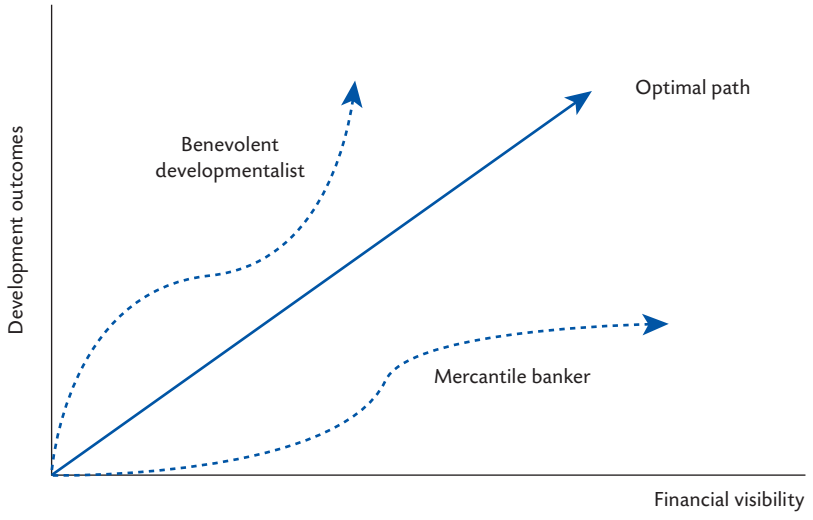
institutional emphasis, especially in undemocratic regimes, on the pretence of national sovereignty. Moreover, there is limited knowledge on how exactly to influence the development of good policies and institutions. Donors know good institutions when they see them, but they know less how to engineer them in a particular country. Furthermore, institutions develop very slowly and in a complex fashion. This is not particularly encouraging for typical aid agencies that are bound by short-term "key performance indicators" tied to short-run results. The lack of patience therefore explains the inadequate investment in institution building and in developing the capacity to implement good policies.

6.3.4. Poor alignment of incentives and interests

Aid effectiveness is also compromised by the lack of convergence between the interests of donors and those of recipients. Within donor governments and development financing institutions (DFIs), there is often also a lack of consistency between the incentives and interests of the institutions and those of operations officers. Like any investment venture, development aid carries risks. Yet, it is by taking risks that aid can generate the highest rewards in terms of development outcomes. Thus, the donor must strike a balance between financial risk and development outcomes (Figure 12). The tendency of programme and project officers is to err on the safe side, minimising financial risks to demonstrate that money has been used well, thus staying on good terms with internal audit. For development financial institutions, this risk aversion is further motivated by the need to preserve the financial bottom line and good credit ratings. As illustrated in Figure 12, donors face tensions between maximising development

Figure 12

Development-risk trade-off



outcomes, or behaving as "benevolent developmentalists", and maximising financial viability, or acting as "mercantilist bankers". In such a context, risk aversion tends to keep aid below the optimal path with regard to development outcomes.

Another way in which incentive incompatibility undermines aid effectiveness is through the pressure to spend the aid budget in the context of the "spend it or lose it" culture of aid budgetisation. This practice induces inefficient allocation of aid resources to activities with little development gains. This causes problems with aid absorption, as large volumes of aid are appropriated to projects and programmes but remain unutilised for long periods. The pressure to "move the money" creates allocative as well as management inefficiencies, especially since programme officers are evaluated not on the basis of effectiveness but on the basis of approvals. Ultimately, these micro level inefficiencies contribute to the overall weak effectiveness of aid at the macro level.

6.3.5. Lack of learning

The challenges described above have been pointed out repeatedly for a long time; yet they continue to permeate the development aid practice. A key reason for this is the imperfection of the learning process in the development aid industry. Evaluation is often not integrated into aid and development policy, and there are inadequate investments in developing evaluation mechanisms. This prevents the development of what Ostrom *et al.* (2001) call the "error-correction capabilities" of systems and institutions that prevent mistakes from generating self-perpetuating inefficiencies. Thus improvements in aid effectiveness are conditional to the development of effective learning. The question this paper raises is whether evaluation can help fill this learning gap and, as a result, contribute to bridging the gaps between macro level and micro level aid effectiveness. This question is the object of the next section.

6.3.6. Inadequate reporting and dissemination of the concrete impacts of aid

Even in situations where aid has been effective and produced tangible results, especially at the micro and sectoral level, often the results remain unknown to the recipients as well as to the public in the donor countries. It is generally observed that Non-Governmental Organisations do a better job in publicising their interventions and drumming up the results of their projects. This somehow explains the relative success of actions by specialised agencies, and it is consistent with the evidence on aid effectiveness at the micro level discussed above.

The lack of transparency and inadequate dissemination of the results of aid arise primarily from the tradition that aid management is the government's domain. In developing countries where institutions of public accountability remain under-developed, government operations are not open to the public, and participatory budgeting is not part of the policy and political culture. Thus, while the population is arguably the ultimate beneficiary of aid interventions, it is not systematically informed of the nature of interventions and their concrete results. It is expected that increased democratic consolidation and the

development of a free press will lead to more pressure on governments to open up the aid management process, which will result in better access to information on aid effectiveness for the general public. This is key to building political support for development aid.

Moreover, the technical process of reporting on aid effectiveness remains inadequate and it is not systematically integrated into the programming and delivery of aid. Even in donor countries where institutions of public accountability are developed, the general public has inadequate access to reports on the results of aid interventions. Reporting to the general public is often defensive, reacting to criticisms from the media and the research community rather than being seen as an inherent obligation of government aid agencies. This tradition undermines the overall aid effectiveness agenda.

6.4. Evaluation can play a major role in bridging the micro-macro gaps

6.4.1. Substantial progress has been made in evaluation practice but challenges still remain

Evaluation is a key component of national aid policy, helping to set goals and access performance. For the USAID (2011: 1), "evaluation is the means through which it can obtain systematic, meaningful feedback about the success and shortcomings of its interventions. Evaluation provides the information and analysis that prevents mistakes from being repeated, and that increases the chance that future investments will yield even more benefits than past investments." Similarly, the Japanese International Cooperation Agency (JICA) considers that "the primary objective of evaluation is to improve the effectiveness and efficiency of projects by using evaluation results for better planning and implementation" (JICA, 2004).

There has been progress in evaluation methods and practice, especially with the introduction of experimental methods in the design and implementation of evaluation. The use of randomised experiments and randomised control trials (RCTs) holds promise; they are scientific, objective, and minimise sampling bias, thus enhancing the reliability of evaluation results. These methods also have the advantage of being replicable in various settings (Duflo and Barnejee, 2009; Duflo and Kramer, 2005). The use of control groups enables the analyst to get closer to establishing a causal relationship between a particular intervention and the targeted outcomes.

But there still are many issues, even with the RCT methodology. In particular, the limitations of RCTs stem from the fact that the method works well in situations where an intervention is truly discrete and homogeneous across space and time (Bamberger and White, 2007). This obviously happens in scientific laboratories, but rarely in real social settings. Moreover, RCTs do not completely overcome the perennial problems of attribution of outcomes in a complex system like an economy, where many factors are likely to influence directly and indirectly a particular outcome (Vaessen, 2010). Furthermore, randomisation may face ethical problems as the exercise involves groups that are not benefiting from an intervention that they would otherwise have wished to benefit from. In such a context, it is difficult to explain why some groups would only serve as experimentation objects while others are beneficiaries of the aid intervention under evaluation.

Regardless of the particular methodology used, evaluation faces structural problems.

Key problems are briefly described here. One of the frequent challenges in evaluation is the lack of a clear definition of instruments and targets as well as limited understanding of the exact channels through which the instruments are expected to generate outcomes. This manifests itself in weak logical frameworks of programmes and projects. So, for example, the logical frame of a road project often lists among expected impacts an increase in GDP growth. Certainly, there are many intermediate targets between road construction and GDP growth, and unless the entire chain of causation is clearly defined, it is difficult for evaluation to be effective. Moreover, evaluation faces the classic problem of discrepancy between instruments and goals when too many goals are set with too few instruments. Thus, when the evaluation fails to find the expected outcome of aid, it is difficult to know whether the failure is due to a bad choice of instruments or bottlenecks in the intermediate causal relationships. This also means that innovations such as the so-called "results-based-management" frameworks cannot be effective without a sound definition of targets, instruments and transmission mechanisms.

Moreover, without a good baseline and control groups, reasonable relative progress may be misjudged as failure. For example, in a post-conflict country, large improvements in institutional and economic performance are difficult to achieve in the short run. To illustrate the point, in these countries, achieving the MDGs to reduce poverty by half by 2015 may be impossible. But there may be substantial improvements relative to the no-project scenario. In this case, without reasonable evaluation criteria, interventions in such settings are inherently set to fail – the "set-to-fail" syndrome. Here the recommendation would be to look at not only the achievement of final targets but also the extent of efforts and relative improvements. To use a sports analogy, evaluation should seek to crown not only most valuable players but also most improved players.

While all donors consider evaluation as an important tool for aid planning and management, they are nevertheless aware of the possible negative repercussions that stem from negative evaluation results. Unsatisfactory evaluations may jeopardise new aid budget allocations (by Congress or Parliament) and even damage the relationships between donor and recipient governments. On the recipient side, there is a risk that negative evaluations may jeopardise new aid. These risks may cause both an underinvestment in evaluation and delayed evaluations due to the "fear-of-the-unknown" effect.

6.4.2. How can evaluation contribute to bridging the micro-macro gap?

Well designed and carefully executed evaluations can help establish better linkages between micro level aid outcomes and the macro level impacts. In other words, to the extent that evaluations are well implemented and systematically integrated along the entire operational cycle, they can help optimise the aggregation of project or programme-level outcomes into national level impacts. This requires a number of innovations in the way evaluations are designed, implemented and

utilised in aid policy. Below are key avenues of possible improvements in that regard.

Evaluation as development diagnostics

Evaluations can help bridge the macro-micro gap to the extent that they are conceived as "development diagnostics" aimed at uncovering the key drivers of intended development outcomes, as well as the channels of transmission from the intervention to the ultimate outcomes. This requires deep knowledge of the sectors involved, the specificities of the country and region, including economic and non-economic features that influence the behaviour of agents and overall economic performance. Evidently such knowledge does not necessarily have to be generated within a particular evaluation exercise. What is needed is close synergies between applied research and evaluation to make evaluation genuinely knowledge-intensive.

Comprehensive ex-ante evaluation as decision-making tool

Many multilateral DFIs have endorsed the practice of *ex-ante* evaluation of development outcomes as a tool for guiding decisions in private sector financing operations. The objective is to identify and attempt to quantify the expected additionality and development outcomes of private sector operations. However, despite the increase in private sector portfolios of DFIs, the dominant lending window remains the public sector. The latter is still not covered by *ex-ante* evaluation of additionality and development outcomes in most DFIs. Two innovations are needed to harness the value added of *ex-ante* evaluations. First, these evaluations need to be extended to the entire portfolio of multilateral financing institutions including public sector operations. Second, *ex-ante* evaluations need to be more comprehensive and address all aspects of development outcomes, including policy and institutional impacts. At the moment, the analysis on institutional impacts and potential implications for policy is still limited. Yet, this is an area of great potential value added towards improving aid effectiveness and bridging the micro-macro gaps in aid effectiveness.

Evaluation to enhance policy and institutional impact of aid

As discussed earlier, a major weak link in the aid effectiveness chain is the limited contribution of aid to improvements in policy and institutions. Part of the reason is that it is typically not an explicit goal of aid interventions to improve policy and institutions. This is either because institutions and policy are seen as too sensitive or simply because it is believed that aid cannot meaningfully influence policy and institutions. Yet for aid to be more effective at the macro level, or for micro level interventions to translate into national development outcomes, it is indispensable to incorporate improvements of policy and institutions as part of the goals of aid. It is therefore important for evaluations to pay particular attention to the impact of aid on policy and institutions. In this way, evaluations can help the aid process by identifying the factors that make aid effective in improving policy and institutions, and by uncovering the mechanisms and channels that generate such positive impacts. This requires a rethinking of the design and implementation of evaluation frameworks to

bolster the policy and institutional dimensions.

Better integration of evaluation outcomes into operations

The evaluation functions are understandably typically separated from the lending functions of most institutions and governments. This preserves the independence of the evaluator to ensure the credibility and reliability of evaluation findings. However, independence carries some costs. It prevents optimal use of the feedback from evaluation in policy formulation and in the design and implementation of operations. Moreover, the feedback from policy design and operations into evaluations is imperfect. Hence the learning is suboptimal, with the risk that errors and mistakes be repeated over time. An evaluation is good only if it informs policy. One way out is to require that programme officers systematically demonstrate that past evaluation has been integrated in the design and implementation of new interventions. Explicit requirement to build upon lessons from past evaluations would promote the institutionalisation of integration of evaluation into operations.

Transparency, participation and public disclosure of information

For evaluations to serve as an effective tool of aid effectiveness, it is essential to develop a culture of transparency, participation and public disclosure of information in donor agencies and the donor community in general. While most DFIs have established policies on disclosure and access to information,[53] these policies are rarely known by the target public and they are poorly implemented partly due to lack of adequate resources. Disclosure of information on aid is even less frequent in many government agencies. The increase in the number of donors is accompanied by growing disparities in the practice of information disclosure on aid, despite calls for donor coordination and harmonisation. Yet, public disclosure of information is important to enable the recipient populations as well as the public in donor countries to keep up with the use of aid resources and their concrete impact on development. Thus, transparency and public disclosure of information are key to the aid effectiveness agenda.

[53] The World Bank Policy on Access to Information dates from 2010; The African Development Bank is at an advanced stage of updating its Policy on Disclosure of Information.

6.5. Conclusion

The debate on aid effectiveness has been disproportionately focused on macro level outcomes that look at the impact of aid on national development outcomes such as growth, improvements in the quality of life brought about by better education, and the health status of the general population. However, individual aid interventions do not actually affect these outcomes directly. While aid effectiveness at the aggregate level remains unsatisfactory, the aid landscape contains individual success stories at the micro level. The dilemma is how to bridge this micro-macro gap.

The analysis in this paper suggests that increasing aid effectiveness will require improvements at three levels: (1) improved effectiveness of aid at the micro-level; i.e. at project and sector level; (2) more transparency, better reporting and public disclosure of information on development outcomes; and (3) better aggregation of micro level outcomes into macro level impacts. The paper argues that such an aggregation technology must be knowledge- and institution-intensive. Institutions are critical for not only the quality of outcomes of individual policy interventions (*i.e.* micro level effectiveness), but also for facilitating positive spillover effects of individual interventions into the rest of the economy. Institutions also facilitate learning from experience, which is essential to improvement in policy design, implementation mechanisms and the overall impact of aid at the macro level.

For evaluations to contribute to bridging the macro-micro gaps in aid effectiveness and help dissipate the clouds surrounding the impact of aid, some innovations in the design and implementation of evaluations, and reporting on aid results are essential. First, this requires a substantive increase in the knowledge intensity of evaluations. Secondly, it is important to achieve higher systematic utilisation of evaluation-generated knowledge in policy and programming than observed in current practice. Third, it is important to improve the reporting mechanisms and systematically embed aid reporting and dissemination of results into the aid planning and delivery processes both at the project and programme levels.

To achieve these innovations, donors and governments in recipient countries need to put their money where their mouth is: if they believe in evaluation, then they must adequately resource it. This requires investing more in evaluation through higher budgetary allocations. It is also necessary to invest more in building capacity and skills in evaluation at the donor and recipient country level. Moreover, it is imperative to develop a culture of transparency, openness and public disclosure of information on aid management in donor and recipient countries. This will improve accountability and ultimately enhance aid effectiveness.

References

Africa Infrastructure Country Diagnostic (2009), *Africa's Infrastructure: A Time for Transformation, World Bank, Washington D.C, available at* www.infrastructureafrica.org

Bamberger, M. and H. White (2007), "Using strong evaluation designs in developing countries, Experience and challenges", *Journal of Multidisciplinary Evaluation,* 4(8): 58–72.

Birdsall, L, N. (2007), "Do no harm, aid, weak institutions and the missing middle in Africa", *Development Policy Review,* 25(5): 575–598.

Burnside, C. and D. Dollar (2000), "Aid, policies, and growth", *American Economic Review,* 90(4): 847–68.

Clemens, M., S. Radelet, and R. Bhavnani (2004), "Counting chickens when they hatch, the short-term effect of aid on growth", Working Paper 44, Center for Global Development, Washington DC. (July).

Collier, P. and J. Dhen (2001), "Aid, shocks, and growth", Working Paper 2688, World Bank, Washington DC (October).

Collier, P., and D. Dollar (2004), "Development effectiveness: What have we learnt?" *The Economic Journal,* 114 (496), F244–71.

Collier, P. and A. Hoeffler (2004), "Aid, policy and growth in post-conflict societies", *European Economic Review,* 48(5): 1125–45.

Dalgaard, C.-J. and H. Hansen (2001), "On aid, growth and good policies", *Journal of Development Studies,* 37(6): 17–41.

Dalgaard, C.-J., H. Hansen and F. Tarp (2004), "On the empirics of foreign aid and growth", *The Economic Journal,* 114(496): F191–216.

Doucouliagos, H. and M. Paldam (2005), "The aid effectiveness literature, the sad results of 40 years of research", Working Paper 2005–15, Department of Economics, University of Aarhus, Aarhus, Denmark.

Dreher, A., P. Nunnenkamp and R. Thiele (2007), "Does aid for education educate children? Evidence from panel data", Mimeo, The Kiel Institute for the World Economy (January).

Duflo, E. and A. Banerjee (2009), "The Experimental approach to development economics", *Annual Review of Economics,* 1: 151-178.

Duflo, E. and M. Kremer (2005), "Use of randomization in the evaluation of development effectiveness", in George Pitman, Osvaldo Feinstein and Gregory Ingram (eds.), *Evaluating Development Effectiveness*, Transaction Publishers, New Brunswick, NJ 2005, pp. 205–232.

Easterly, W. (2006*a*), "The Big Push Déjà Vu, A Review of Jeffrey Sachs's *The End of Poverty, Economic Possibilities for Our Time", Journal of Economic Literature,* 44 (March): 96–105.

Easterly, W. (2006*b*), *The White Man's Burden, Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*, Penguin Press, New York.

Easterly, W. (2009), "Can the West Save Africa?", *Journal of Economic Literature,* 47(2): 373–447.

Easterly, W., R. Levine and D. Roodman (2004), "Aid, policies, and growth: Comment," *American Economic Review,* 94(3): 774–80.

Guillaumont, P. (2007), « La vulnérabilité économique, défi persistant a la croissance africaine », *Revue Africaine de Développement,* 19(1): 123–162.

Guillaumont, P. (2009), "An economic vulnerability index, its design and use for international development policy", *Oxford Development Studies,* 37(3): 193–228.

Guillaumont, P. (2010), "Assessing the economic vulnerability of small island developing states and the least developed countries", *Journal of Development Studies,* 46(5): 828–854.

Guillaumont, P. and L. Chauvet (2001), "Aid and performance, a reassessment", *Journal of Development Studies,* 37(6): 66–92.

Guillaumont, P. and S. Guillaumont-Jeanneney (2009), "Accounting for vulnerability of African countries in performance based aid allocation", African Development Bank Working Paper No.103 (October).

Guillaumont, P. and C. Simonet (2011), "To what extent are African countries vulnerable to climate change? Lessons of a new indicator of physical vulnerability to climate change", CERDI / Université d'Auvergne.

Hansen, H. and F. Tarp (2000), "Aid effectiveness disputed", *Journal of International Development,* 12(3): 375–98.

Hansen, H. and F. Tarp (2001), "Aid and growth regressions", *Journal of Development Economics,* 64(2): 547–70.

Independent Evaluation Group (2008), "IFC's Experience and Additionality in Middle Income Countries Results and Challenges," Note No.1.

IFC (2010), "IFC's Additionality: A Primer", September 30, 2010.

JICA (2004), JICA Guideline for Project Evaluation – Practical Methods for Project Evaluation. (September).

Kanbur, R. (2000), "Aid conditionality and debt in Africa", In F. Tarp (ed.) *Foreign Aid and Development – Lessons Learnt and Directions for the Future*, Routledge, London, pp. 409–422.

Killick, T. (1998), *Aid and the Political Economy of Policy Change*, Routledge/ODI, London.

McGillivray, M., M. S. Feeney, N. Hermes and R. Lensink (2005), "It works; it doesn't; it can, but that depends..., 50 years of controversy over the macroeconomic impact of development aid", Working Paper 2005/24, World Institute for Development Economics Research, Helsinki (August).

MDGs Africa Steering Group (2008), "Achieving MDGs in Africa", New York, United Nations.

Michaelowa, K. and A. Weber (2006), "Aid effectiveness reconsidered, panel data evidence for the education sector", Discussion Paper 264, Hamburg Institute of International Economics.

Mishra, P. and D. Newhouse (2007), "Health aid and infant mortality," Working Paper 07/100, International Monetary Fund, Washington, DC. (April),

Mold, A. (2009), *Policy Ownership and Aid Conditionality in Light of the Financial Crisis – A Critical Review,* OECD, Paris.

Mosley, P. (1980), "Aid, savings and growth revisited," *Oxford Bulletin of Economics and Statistics,* 42(2), 79–96.

Ostrom, E., C. Gibson, S. Shivakumar and K. Andersson (2001), *Aid, Incentives, and Sustainability, An Institutional Analysis of Development Cooperation*. Sida Studies in Evaluation 02/01.

Clemens, M., S. Radelet, and R. Bhavnani (2004), "Counting Chickens When They Hatch: The Short-Term Effect of Aid on Growth", CGD Working Paper No. 44, the Center for Global Development,

Washington DC Rajan, R. and A. Subramanian (2005), "What Undermines Aid's Impact on Growth?", IMF Working Paper 05/126, International Monetary Fund, Washington.

Roodman, D. (2007*a*), "Macro Aid Effectiveness Research: A Guide for the Perplexed", Working Paper 134, Center for Global Development.

Roodman, D. (2007*b*), "The anarchy of numbers: aid, development, and cross-country empirics", *World Bank Economic Review,* 21(2): 255–77.

Sachs, J. (2005), *The End of Poverty, Economic Possibilities for our Time*, Penguin Press, New York.

Sachs, J., J. McArthur, G. Schmidt-Traub, K. Kruk, C. Bahadur, M. Faye and G. McCord. (2004), "Ending Africa's poverty trap", *Brookings Papers on Economic Activity*, 1.

Svensson, J. (1999), "Aid, growth and democracy", *Economics and Politics,* 11(3): 275–97.

Unitted Nations MDG Task Force (2011), *Millennium Development Goal 8 - The Global Partnership for Development, Time to Deliver*, United Nations, New York.

USAID (2011), Evaluation – Learning from Experience. USAID Evaluation Policy. (January)

Vaessen, J. (2010), "Challenges in impact evaluation of development interventions, opportunities & limitations", Discussion Paper 2010/01, Institute of Development Policy and Management, University of Antwerp.

7. Applying Evaluation to Development Policy

Miguel Székely, Institute for Innovation in Education
Tecnológico de Monterrey

Abstract

This paper addresses two concerns that have gained attention in the development and social policy literature in recent years: on the one hand, it is perceived that the amount of evidence and knowledge on the effects on different types of policy interventions on development results is still scarce. On the other, there are also documented perceptions that the available evidence is being used to a very limited extent for improving policy. We argue that some of the reasons for this apparent paradox are: (i) the limited scope and information that can be provided by evaluations; (ii) the differences in the incentives and objectives of the different actors involved in the evaluation generation and use process; and (iii) the differences in capabilities to use and produce information for improving policy design. We argue that in order to move forward it is necessary to examine these elements and go beyond looking separately at the supply and demand for evidence, which appears to be currently the prevalent view, and visualize more integrated approaches.

7.1. Introduction

This paper addresses two concerns that have gained attention in the development and social policy literature in recent years. On the one hand, the perception is that evidence and knowledge on the effects of different types of policy interventions on development results are still scarce.[54] On the other hand, there are documented perceptions that the available evidence is being used to only a very limited extent to improve policy.[55]

So why is knowledge and evidence on what works for human development scarce and at the same time under-utilised? One reason relates to the scope and type of information that can be provided by evaluations. As is generally the case in all social sciences, taking human behaviour as an object of analysis to help evaluate the impact of a particular action involves a high degree of uncertainty, and is much more complex than measuring expenditures or inputs. While most expenditures and inputs can be measured and monitored through auditing and accounting mechanisms, outcomes require an understanding of individual behaviour and reactions. Even though substantial methodological advances have been made over recent decades, the literature on this subject can still only offer approximations when assessing the full impact of policy actions.

Another central element concerns the incentives and objectives of the different actors involved in the production and use of evaluations. Yet another concerns the capacity to produce and use and produce information to improve policy design. We argue that in order to move forward it is necessary to examine these elements and go beyond the currently prevalent approach of examining separately the supply and demand for evidence, and to envisage more integrated approaches.

We develop our argument in three sections. Section 7.2 examines the behaviour of the main actors involved in the process of producing and using evaluations for development effectiveness. Section 7.3 presents some evidence from Mexico that is a useful reference point for identifying the elements of an evaluation system that could reconcile supply and demand concerns. Section 7.4 concludes by pinpointing some of the components that could be included in a more integrated system aimed at improving policy.

[54] See for instance Savedoff, Levine and Birdsall (2006).

[55] Examples include Ravallion (2008), Jones et al. (2009) and Weiss (1999).

7.2. Understanding the actors involved in the process

In order to understand the limited production and use of evaluation in development policy, one necessary step is to analyse the motives, context and profile of each of the actors involved in the processes. At least five different actors with varying priorities, interests and preferences normally coexist to generate the outcomes currently observed:

(a) external donors who allocate resources to countries, programmes or specific actions, and who may demand evidence on the impact of the activities carried out using these resources;

(b) high-level policy makers, who, on the one hand, are responsible for improving policy at the general or sectoral level and, on the other, have the power to authorise and mandate the evaluations;

(c) evaluators responsible for performing the analysis, and who may have their own interests and motives;

(d) practitioners or programme operators/executors who implement the programme's actions in the field or deliver a service; and

(e) constituencies, public opinion, direct/indirect beneficiaries and other actors that call for the effective use of public resources.

Below, we examine some of their motives for producing or using evaluation results, as well as the limitations and circumstances that may shape their behaviour. We suggest that the differences in objectives, incentives and motives of the various actors are determining factors of the reality we observe today.

7.2.1. External donors, investors

Among the most influential actors in evaluation production and use are donors or investors that are external to the operation of the programme and to the agency in charge of its execution, but who provide financing for the implementation of policies and actions. They range from multilateral institutions and foundations to local-level private actors that are interested in supporting efforts to address certain problems. There are several reasons why these actors can deliberately choose not to advocate measuring the impact of their resources on development results and, in some circumstances, they may even intentionally undermine efforts to evaluate it. The first reason is that evaluation requires resources. Normally, for proper evaluation, tailor-made data must be produced, resources must be invested in developing an appropriate evaluation design, and funding needs to be found for the professionals who analyse the data and report the findings and conclusions. When resources are limited

(which is most usually the case), a moral dilemma arises as to whether funding should be allocated to evaluating impact, or to benefiting more people. The obvious argument in favour of undertaking an evaluation is that it will help to verify whether the intervention has any positive or negative impacts, and that its results will help to enhance the effectiveness and efficiency of resource spending, with further benefits for larger populations in the future. The argument against devoting resources to this activity is that when needs are critical, diverting funding away from a beneficiary might make the difference between a human being's life and death.

A second reason is that since knowledge is a public good, the incentives to allocate resources to evaluation, whose benefits may stretch well beyond the immediate interest of learning more about an intervention's impact, declines. This argument is discussed in detail by Avery *et al*. (1999) and Savedoff *et al*. (2006). This public-good aspect of evaluation is exacerbated when multiple donors participate in fund-raising for large-scale initiatives, as each donor perceives their contribution as no more than marginal (a drop in the ocean). This perception means that each donor's incentives and empowerment to demand accountability are particularly weak.

A third reason is that, in some circumstances, the real underlying objective of external donors/investors might not be to generate impacts, but rather to make the statement that they are supporting a particular cause. In these cases, the objective may well be a noble and legitimate one, and the measure of success will be the actual flow of resources rather than the final impact of an intervention, but ultimately this does not incite either the donor/investor or the executor to invest in evaluation.

A fourth reason, also related to the economic cost of performing evaluations, is that private donations may be used for self-serving goals including commercial interests or tax exemptions. In some cases, donations may be attractive not only for their development impact as such, but also because they offer an alternative to paying taxes (e.g. individuals or organisations that may prefer any option other than providing funding for the government), or they possibly offer an opportunity to promote or advertise commercial products. In this scenario, there are fewer incentives to evaluate impact.

Evidently, there are also strong reasons why external donors/investors genuinely interested in promoting development can advocate and even fund evaluations. Perhaps the strongest reason centres around the principal-agent problem, which is to say that unless evidence is produced on the use of the resources by a credible (usually external) party, the principal (donor/investor) cannot ensure that its funding is being used for the intended purposes, since the agent (receiving agency/executor) might have their own different priorities and preferences. This is discussed in detail in Birdsall and Savedoff (2010), who propose the "cash-on-delivery" mechanism as a way of ensuring that investment in aid and other kinds of support yield development results. Another reason is that external agents with a longer-term perspective might find it profitable to invest in evaluation in order to produce the

necessary knowledge to guarantee greater effectiveness and efficiency in their own future investments. Yet another important explanation is that, when external donations and investments depend on fund-raising, the evidence from evaluations may be critical in convincing donors to continue their efforts. The lack of such evidence might be detrimental to the action's sustainability.

In summary, the attractiveness of performing evaluations is not self-evident from the standpoint of the external donors/investors that finance public social programmes and actions. In cases where their motive involves the resource flow *per se*, evaluation may even be viewed as no more than a financial burden. The lack of incentive to support evaluation in these cases will most probably trickle down to the other actors involved, resulting in low or no investment in this activity. On the contrary, when the motive is to achieve development results, strong incentives will normally exist to carry out evaluation.

7.2.2. High-level decision-makers

The second strategic agents are the high-level policy- and decision-makers, who may be at the same time (indirect) producers and users of evaluation results. They are potential users in the sense that general decisions of policy orientation, funding and implementation are under their responsibility, and the information generated by evaluations can be strategic for identifying areas for improvement and providing feedback for better decision-making. They can also be considered producers in the sense that they are in charge of authorising the evaluation of a public programme or activity in their domain. In this case also, we could mention reasons for or against authorising and devoting resources to evaluations.

Even from the limited perspective of optimising the time spent in an influential government position, information from evaluation may be a powerful tool insofar as it allows positive results to be communicated to one's constituencies, thus bolstering their support. In some cases, it may have the advantage of offering insights on how to achieve even better future results (and thus greater support). One explanation for the limited use of evaluation, despite these potential advantages, is that policy decisions are taken within a set of constraints and conditions that may overshadow the benefits of evaluation. Some of the most common constraints include timing, interest groups, normative factors, institutional decisions, politics and funding. The question of timing is often at odds with the dynamics of producing credible evidence. In public actions, time frames for delivering results are usually extremely tight, whereas a robust impact assessment needs time to develop the evaluation design, to produce baseline data, to allow the intervention to produce its effects, to generate *ex-post* information and finally to analyse and report results. All of this could imply years of investment that stretch far beyond the political cycle, which could well discourage even the most enthusiastic proponents of evaluation as they would not be able to reap the benefits of their initiatives.

Interest groups may also represent a considerable constraint for high-level policy makers, as such groups may dissuade them from carrying out impact evaluations,

especially if these involve experimental designs. For example, the definition of control and treatment groups might be perfectly justifiable from a methodological point of view, but it is no easy task to explain to the non-participants of a programme that they have been excluded from a benefit because they were not "randomly selected", while others have been selected. This may create enough opposition to make the evaluation unfeasible. Even in cases where an experiment can be launched, the evidence of positive results may understandably generate pressure from the control group to be integrated into the programme, thus creating the risk of "contaminating" the experiment. Interestingly, these pressures usually arise when the control group has to be maintained intact over considerable period of time, since the full effects targeted can only be measured when the intervention has time to yield its medium- or long-term impacts. The policy makers' capacity to deal with political and interest group pressures usually determines the feasibility of the experiment.

Normative factors may also play an important role. If, for instance, rules and regulations exclude using programme resources to evaluate impact, even if the policy makers are proponents of evaluation, there will be underinvestment in this activity. Similarly, institutional arrangements may impede evaluation. When government institutions are designed around the concept of measuring expenditures and inputs, the mandate to evaluate outcomes or implementing agencies may be inexistent, making it impossible or even illegal to perform evaluations.

And there is also the political side of things. Producing information on the efficiency of policy action may be highly risky in some settings. While positive effects can be capitalised politically, unfavourable results may be much more difficult to handle in certain circumstances and managing them may require investing political capital. As argued by Pritchett (2002), the risk of obtaining negative or not-so-positive results from evaluation might be a strong deterrent to promoting it. Results may provide the opposition with ammunition, they may backfire and become politically lethal or discourage external donors/investors from allocating additional resources. The risk is usually higher in societies where strong transparency and accountability mechanisms are institutionalised, and in environments with tight budget constraints (where, paradoxically, there is a greater need for information on which policies are more effective), in which many interest groups are competing for resources. Providing sound evidence on programme impact in such contexts may be equivalent to signing its death certificate.

In the end, the balance between the restrictions and circumstances described above will be critical in establishing the feasibility of producing and using evaluation from the point of view of high-level policy-makers. An important determinant of which way the balance will go is the technical skill required to understand evaluations and assimilate their results. Weak professional capacities at this level may thus increase the risks of evaluating government action, and therefore reduce their attractiveness.

7.2.3 Evaluators

Evaluators also play an important role in the process. These actors are frequently external to the operation of the programme and have a sufficient degree of independence to guarantee credibility. Self-evaluations can also be performed, but as they are prone to subjectivity, this impairs their credibility. As recognised at least since Alkin and Daillak (1979), the approach chosen by evaluators is a critical factor in either promoting or discouraging the production and use of evidence. On the positive side, generating sound and credible information on the effect of policy action may be of high value for external donors/investors, who need assurance on the development effect of their intervention. For policy makers, timely, credible and relevant information that helps capitalise positive government action and offers possibilities to improve policy performance will also be appreciated. The political risks of not-so-good results can sometimes even be mitigated if the policy maker and programme operators intervene early on in the process to help define the questions to be answered, as well as the strategy for adequately communicating results.

However, choosing approaches that restrict the use of evaluations may dampen the interest shown by policy makers and/or external donors and investors. For instance, an exercise that requires waiting a whole generation before information can be obtained may be totally unfeasible for both. Similarly, methodological approaches that generate results that are only valid for a particular setting may be of little use for scaling up or informing decisions in other contexts. Too great a focus on academically interesting but operationally irrelevant issues will also deter the demand for assessments. Also, findings that are methodologically sound but politically unviable will usually encounter a less enthusiastic reception, while evaluation designs that are technically excellent but require an unrealistic budget or are politically unmanageable will most likely not be undertaken – or if they are, will most likely be seldom used.

These situations may arise because the objectives and constraints of the above-described external donors/investors and policy makers do not necessarily coincide with those of the professionals capable of performing robust evaluations. Evaluators (mostly from academic circles) may prioritise academic purity, professional prestige, recognition, knowledge production, academic success (in terms of high-profile publications), etc. These priorities may be inconsistent with evaluations that need to be timely, credible, relevant, pertinent and communicable from the users' perspective. Such incompatibility with user needs may be an important factor in explaining the perception that available evidence is underutilised.

7.2.4. Programme operators and practitioners

For several reasons, the actors perhaps most affected in practical terms by the process of generating evidence from evaluations and implementing changes to programmes and services in line with their results are the programme operators and practitioners working directly in the field. To begin with, from the perspective of producing evidence, coming up with new types of data and

information in order to design and implement assessments usually requires additional work and efforts that are not necessarily accompanied by identifiable short-term benefits or incentives for these actors. Additionally, failure to deliver benefits to specific populations within strict schedules may jeopardise an evaluation or contaminate an experiment, with serious consequences for the quality of the exercise. Yet another reason is their direct exposure to pressure from individuals in the control or other groups, who wish to be included in the programme, in the case of experimental designs.

But perhaps the foremost challenge has to do with the use of evaluation results. Shifting from a policy approach where expenditures or inputs are the yardstick of success towards an approach that defines development outcomes as the benchmark is in itself an important cultural change. When new evaluations are presented and a set of recommendations that modify day-to-day practices, procedures, norms and processes are introduced, it is precisely these field operators that "bear the brunt" of implementing such changes. Phasing new standards and methods into day-to-day operations can be the most complex and laborious task if evaluations are to be used effectively, since it necessarily requires a change in individual behaviour and practice. Even seemingly slight changes in procedures and norms most often take several years to implement fully, and explicit or implicit opposition at this level may inhibit evaluation use.

In some settings, it is also common for operators and practitioners to become "constituencies" of the programme to which they have devoted years of effort and experience. When operators become clients of their own programme, they can be the first to obstruct or openly oppose change and make evaluation use effectively impossible in practice. Even if there is strong commitment at the higher decision-making level to use evaluation results to improve policy design and implementation, resistance at the grassroots level may make this impossible. This may be either because it practically requires additional effort, or because operators have become so closely identified with the programme and the way that it is designed and operated that they take any challenge to the *status quo* personally.

Intensive training and communication on the nature and purposes of evaluations may help ameliorate negative opposition to their production and use. Experience shows that involving these actors in the design process (for instance, by having them participate in the definition of the questions addressed by the evaluation), including them in the hypothesis-setting process, and even having them contribute to the identification of potential areas for improvement may help refocus their efforts towards a more effective use of the evaluation results.

7.2.5. Constituencies, public opinion and beneficiaries

Political constituencies, beneficiaries and public opinion are also relevant in the process of producing and using evaluations in social policy. As mentioned in the introduction, evaluation is a powerful tool for producing information, making assessments on policy performance and

providing feedback to improve policy action, and these actors play an important role in each of these functions. Making information available on policy impact enhances transparency and allows the public to know how their taxes are being spent, or whether certain goals or benchmarks have been achieved. It also allows for value judgments on whether performance is adequate or not, and therefore serves as a tool for rendering policy makers and programme operators accountable for their actions. Transparency and accountability are highly valued in many political settings and constitute strong incentives to demand evaluation production. In fact, there is generally a positive relationship between the demand for evaluation and the level of transparency and accountability in a society.

Evaluation results can also be used by these actors to promote changes to improve policy effectiveness, although the direct channels through which they can exert an influence on this aspect are not always available.

The ideal setting in which evaluation can fulfil its function as a strategic tool for development is when an informed society that uses evaluation results to demand policy improvements coexists with a receptive, transparent and accountable government that implements improvements and informs and justifies the use of results and recommendations. However, if either side (constituencies or governments) fails to play this role, evaluation may become a threat or even have negative effects. When, for instance, rather than using results constructively to improve policy, constituencies, beneficiaries and public opinion use the evidence simply to expose failure or signal and discredit specific participants, sometimes punitively, there will be a heightened perception that evaluation poses a risk for both operators and policy makers. This type of situation is commonly found in countries where transparency and accountability are novel phenomena, introduced after long periods of censorship or limited citizenship rights. And the longer such periods last, the greater the costs associated with evaluation will be for the public sector. The media usually play a key role in influencing the direction that the discussion and debate take in this regard. Constructive media coverage focused on improvement may lead the discussion towards a follow-up on the use of results, while an extremely critical media focused only on pointing up failure and signalling or penalising the actors involved in policymaking and operations may not necessarily lead to better performance.

In sum, each of the five relevant actors involved in the production and use of evaluation in social policy play an important role individually. Moreover, the interactions among them determine the final outcome. For instance, the combination of extremely critical constituencies and public opinion may inhibit evaluation practices when policy-makers have low technical capacities for using and interpreting evaluation results, or may provoke extreme reactions by programme operators, who could then obstruct further assessments in the future. Similarly, when external donors/investors and policy makers centred on improving policy act alongside professional operators focused on generating outcomes, with feasible and useful evaluation design and implementation, and well-informed, critical,

but nonetheless constructive constituencies, this combination may create a virtuous cycle of knowledge production and policy improvement. As discussed below, the main challenge here is to align the incentives and objectives of all five actors and provide an adequate institutional setting in order to shift away from producing scattered evaluations that are used only intermittently towards a comprehensive evaluation system that promotes continuous improvement.

7.3. Some relevant experiences

To outline some of the main elements of an evaluation system that effectively promotes improved design and implementation of social policy, it is useful to document some recent experiences that highlight the critical components likely to trigger the virtuous circle mentioned above. Some of the elements in this section are drawn from the author's experience in Mexico over the past ten years. Mexico has made important strides, moving from a pre-2000 system based on measuring expenditures and inputs to one of the most developed evaluation systems in Latin America today.

The experiences refer to five practical cases where evaluation instruments, practices and institutions were introduced in a short time span.[56] They include the introduction of evaluation as a general practice in poverty alleviation policy, the definition and implementation of poverty measures, the creation of an evaluation system at the school level, the introduction of a national academic assessment test at the high school level, and the creation of the National Council for the Evaluation of Social Policy (CONEVAL). From each of these experiences, we can gain important insights into the design of a more comprehensive approach that aligns incentives and objectives of the five actors discussed in the previous section.

7.3.1. The evaluation programme at the Ministry of Social Development: evaluating one programme is useful, but not enough

In 1998 Mexico launched the PROGRESA conditional cash transfer programme, which, among several other innovations, featured an impact evaluation designed and implemented from the outset. In 2000, evidence of its positive impact on education, health and nutritional outcomes had already been gathered. This played a key role in enabling it to survive as the main poverty alleviation programme despite a change of government that same year (in fact, the only change the new government brought to the programme was to rename it "*Oportunidades*"). Furthermore, the evaluation exercise inspired many others in Latin America and other regions. Having a high-quality external evaluation to hand, especially following decades of policy without research, was to make a huge difference. For the first time, evidence of the impact of each budget unit became available. This not only provides reassurance that resources are being used adequately, but it also becomes the main argument for scaling up and increasing investment.

[56] The examples are cases where the author personally played an active role. Since no first-hand example is available to illustrate the role played by external donors or investors, we are not able to include an illustration for this case.

Once the PROGRESA-*Oportunidades* results became available and were widely used over several years to identify programme improvement areas, the obvious question arose as to whether this was really the best possible use of public resources. To answer the question, the Government decided to introduce similar evaluation designs for other social programmes and it was when more information on their impact became available that better decisions could be taken. This Mexican experience illustrates that, after the shock of shifting to a new culture of evidence-based decision-making in social policy, evaluation practices eventually start to become assimilated, and although resistance may persist – especially in the field – the actors involved at some point internalise the new culture and procedures.

7.3.2. The introduction of poverty measurement: evaluation requires investing political capital

The second case is the introduction of official poverty measurement in Mexico. When the Mexican opposition party came to power in 2000 for the first time in decades, a new window also opened up for the first-ever production of official poverty statistics. One important motive behind this was to document the country's social conditions after so many years of a single-party system. In this case, the two main actors were the high-level decision-makers promoting the setting up of a poverty measurement system on the one hand and evaluators on the other. The Ministry of Social Development took the decision to announce official poverty statistics in the shortest possible time – close on the tail of the new Administration's accession to power – and it invited a group of respected researchers to propose a methodology robust and rigorous enough to ensure credibility and restrict the debate to how this information was to be used rather than on measurement issues.

The official estimate based on the official methodology was that 53.8% of the Mexican population was poor in the year 2000. The perception of unacceptably high poverty rates triggered an unprecedented public debate on national performance, the costs of a non-democratic system, the economic model followed in recent years, etc. The media played a key role in fuelling the debate and it became by far the most discussed social policy issue in many years. Although the data clearly represented the situation before the start of the new Administration, substantial political capital had to be invested in supporting the exercise and guaranteeing its continuation. The government's critics attributed responsibility for the high poverty rates to the current authorities, and opposition parties (mainly members of the party that had ruled the country during the previous decades) made great efforts to discredit the figures, arguing that they were politically manipulated.

The publication of the 2002 figures was less controversial and, after four rounds of official measurements in 2004, 2005, 2006 and 2008, the publication of poverty rates, as well as those for inflation, employment, wages, *etc.*, has become a much less politicised activity and is systematically used to assess the performance of the country and its government. Most importantly, it has

provided a broad framework within which the individual impact of specific programmes can be viewed in a wider perspective.

7.3.3. The planning and evaluation system for schools: time, training, capacity building and resources to implement change are fundamental to producing evaluations that can be used

The third case refers to the decision by Mexico's Education Ministry in 2008 to introduce a new system whereby each school principal (at the high school level) was provided with a spreadsheet in which they were required to enter general administrative data. This spreadsheet automatically generated fifteen selected indicators regarding the infrastructure, materials, equipment, alumni and their own teaching staff. Principals were required to perform two exercises. The first involved prioritising the fifteen variables according to how important each was for improving the quality of education in their specific context and circumstances. The second involved setting a target to improve each indicator during the course of the school year. Starting from a situation of total absence of school-generated data, the main purpose of this activity was to produce a diagnosis (baseline) of school conditions as well as explicit targets, in order to design an annual school improvement plan. In the context of our discussion, the example is relevant as it constitutes a case in which actors critical for the system's operation (programme operators and practitioners) are required to generate inputs for evaluation and at the same time use them to take specific action for improvement.

This apparently simple exercise of setting priorities and targets was not accompanied by an adequate training, information and capacity-building process as it was assumed that school principals already had the necessary managerial skills to perform what seemed to be a simple basic task. After the first year of implementation of the evaluation and planning system, the principals were classified into four different categories. The first included a minority of 3% who had actually completed a high-quality diagnosis and set useful targets to develop an action plan for improvement during the academic year. The second group comprised a substantial 17% of principals who did not perform the exercise at all. The third group, accounting for 52% and made up of the most experienced and older principals, engaged in a more professional process in priority setting but deliberately defined extremely low targets in the belief that low targets were going to be much easier to achieve. The fourth group represented 30% of the total, and included younger and less experienced profiles. This group all deliberately set extremely high targets with the idea that this would give the central authorities the impression that they were ambitious and dynamic.

Following the use of the system over a first academic year and the renewal of leadership in the 17% of schools that had not participated in the first round, some interesting dynamics were observed: the third and fourth groups slowly converged towards setting more realistic and meaningful targets and, most importantly, using the system as a planning device to identify areas for improvement and to demand support from central authorities in

specific areas. One point to emerge from this case is that, once school principals are engaged in the diagnosis-design-benchmarking-implementation-evaluation cycle, resources for introducing improvements can become an important bottleneck. Even when the cultural shift to a knowledge-based system has been accomplished, if adequate resources are not made available to implement improvements, the system can lead to greater stress and frustration rather than creating a virtuous cycle that leads to higher quality. This situation might render the use of evaluation unfeasible in reality and make any future production of evidence irrelevant.

7.3.4. Evaluating educational attainment at the high school level: in order to use information, it is necessary to understand it and build the capacities required to exploit it

The fourth example also refers to education. In 2008, the Federal Government launched national examinations for twelfth grade (exit from high school), with more than 96% of schools participating. As in other cases where prior information is scarce, the first publication of results triggered a huge reaction in the media and public opinion, which rightly criticised the low levels of achievement and identified and aggressively attacked under-performing schools. The natural reaction of the schools (mainly the low performers) was to discredit the test. Strong opposition also emerged from the teacher's union, which felt aggravated by the exposure and criticism.

The main feature of this process is that, even though low-performing high schools have been under intense public pressure to improve their results, after three rounds of application in 2008, 2009 and 2010, their capacity to absorb and internalise the information to take action for improvement has seemed extremely slow. Some schools have chosen to go beyond criticising the test to openly oppose its application, and after a field investigation by the author (of ten non-representative schools in Mexico City), the common complaint across the board was that, while schools were provided with an initial diagnosis and recurrent evaluations thereafter, they were not provided with guidance, orientation or resources in order to introduce improvements and perform better in the next round.

This is a case where evidence is substantial and low use is not due to lack of interest but rather to the lack of local capacities to transform information into better practices. If we were to judge the low levels of production and use of evaluation by this case, the conclusion would be that additional assessments would be welcome, but that until local capabilities at the level of the service provider are strengthened, this valuable information will be permanently under-utilised.

7.3.5. The creation of the National Council for the Evaluation of Social Programs (CONEVAL): the need to institutionalise cultural change and create the right incentives

In Mexico, after taking a first step by creating an Evaluation Programme in the Social Development Ministry, Congress introduced the legal mandate to enforce evaluation of all social programmes funded by public

resources. The broad notion of "social" adopted for this initiative included poverty alleviation, health, education, agriculture, the environment, micro-enterprise and other related sectors. Simultaneously, a Social Development Law was approved in 2004, which included the formal creation of CONEVAL.

The underlying aim was that CONEVAL, which is under the supervision of the Social Development Ministry but independent of its bureaucracy, be allocated the funds to evaluate all social programmes, define a basic structure for their design – for instance, prioritising experimental impact evaluation where possible – and launch requests for proposals among academic, public and private institutions to carry out the evaluations under certain guidelines. This, in addition to the mandate by Congress, guarantees that all programmes will be evaluated regardless of the interests or profile of the decision-makers in charge and of the operators delivering the goods and services. One important feature is the requirement for the programme under evaluation to engage in the evaluation design, for example, by suggesting relevant questions and even by providing feedback on design aspects.

So far, the main impact of CONEVAL and its evaluations has been on transparency and accountability. All evaluations are made public, and since 2006 – which was the first year of formal operation beyond the Social Development Ministry – their presentation has caused intense debate with the media frequently criticising and discrediting government action. The publication process has often created tension and confrontation with other government departments in charge of various programmes, especially as the media still tends to highlight the negative aspects of evaluations and ignore their positive impacts and achievements. Tensions reached a peak when CONEVAL – also in charge of publishing the official poverty statistics since 2005 – released soaring poverty figures for 2008. The institution had to steer through complex situations and questioning from the government authorities themselves, since the news carried high political costs. The fact that it was created by Congressional mandate and is backed by legislation (through the Social Development Law) has proved to be the key asset enabling the institution to maintain its integrity.

After five years of operation, the main challenge is now to go beyond the transparency and accountability benefits and ensure that results are used to actually improve policy. Congress mandates the evaluation of all programmes but has so far not taken a stand on how this information is to be used. In 2010, the elaboration of the 2011 State budget, which by law is defined by Congress, was guided by inertial elements and political arrangements rather than by policy effectiveness and efficiency. As long as budget decisions are not tightly linked to evaluation results, it is unlikely that the social sector in Mexico will fully reach the stage of knowledge-based policy.

7.4. Towards more integrated systems

We propose that a desirable next step in developing countries is to move towards more complete systems – as opposed to only focusing on individual evaluation efforts – that provide incentives for production and use of evaluations, with the support of an adequate institutional setting. To achieve this and based on the experiences described in the previous section, at least four key components would seem necessary. The first involves convincing decision-makers and programme operators. Their role is to set targets, become involved in evaluation design, implement programmes and actions, and use evaluation results to improve the performance of their activities.

The second is the creation of an institution similar in spirit to CONEVAL, which would focus on four activities: defining methodology and evaluation approaches in coordination with the first actor; putting out requests for proposals and selecting evaluators; monitoring the quality of each evaluation; and analysing evaluations to produce reports that identify specific areas and actions for improvement, with responsibility for these being assigned to the first actor.

The third component would be an additional (new) public institution totally focused on capacity building, training and coaching potential users to ensure that results can be translated into action. It seems desirable to disconnect this activity from the design responsibilities (carried out by the second actor) to avoid conflicts of interest. Countries with a strong civil service tradition may already operate along these lines by providing training to civil servants prior to or during their time of service. However, for those not in this case, a dedicated institution ensuring this function may be a more realistic target than waiting for them to develop a full civil service career.

The fourth is an entity that determines public budgets. In the case of political systems where Congress plays this role, the implementation of a more complete system would require a strict discipline that links evaluation results to future funding. There will evidently be other important factors that need to be taken into account for steering budgeting decisions, so an initial step could be to determine a minimum share of the budget to be earmarked for this scheme.

References

Alkin, M.C. and R.H. Daillak (1979), "A Study of Evaluation Utilization", *Educational Evaluation and Policy Analysis*, 1(4): 41–49, July–August.

Avery, C., P. Resnick, and R. Zeckhauser (1999), "The Market for Evaluations", *The American Economic Review*, 89(3): 564–584, American Economic Association.

Birdsall, N. and W. Savedoff (2010), "Cash on Delivery: A New Approach to Foreign Aid", Center for Global Development.

Jones, N., H. Jones, L. Steer and A. Data (2009), "Improving impact evaluation production and use", Overseas Development Institute.

Pritchett, L. (2002), "It Pays to be Ignorant", *Journal of Economic Policy Reform*, 5(4): 251–269.

Ravallion, M. (2008), *Evaluation in the Practice of Development,* World Bank.

Savedoff, W., R. Levine and N. Birdsall (2006), "When will we ever learn? Improving lives through impact evaluation", the Evaluation Gap Working Group, Center for Global Development, Washington, D.C.

8. The Collision of Development Goals and Impact Evaluation[57]

Michael A. Clemens, Center for Global Development

Abstract

Two movements have recently reshaped development aid. The Goal Movement has unified and inspired aid actors with quantified targets; the Evaluation Movement has raised standards for measuring the aid's true effects. These two movements can complement each other, but in some aid projects they have instead unproductively collided. I review one such collision, in the United Nations-sponsored Millennium Villages Project. The story offers lessons on how new development goals and future impact evaluations could do more to reinforce one another.

[57] I am grateful to André Corrêa d'Almeida, Gabriel Demombynes, Charles Kenny, Sara Minard, Jean-David Naudet, Robert Peccoud, and Michael Woolcock for helpful conversations. The views in this paper are strictly my own and do not represent those of the Center for Global Development, its Board, or its funders.

8.1. Introduction

Two social movements have reshaped development aid from within over the past 15 years. The first is the Goal Movement, an attempt by aid policy makers to unify their efforts around measurable declines in poverty by a fixed date. The second is the Evaluation Movement, an attempt by aid researchers to more reliably measure the poverty impacts of aid interventions.

The two movements appear at first to complement each other: both, in some way, emphasise results over process, outcomes over inputs. But many years into the Goal Movement, only a slim fraction of all aid projects receive any rigorous impact evaluation – that is, any reliable assessment of how results with the project were different from what they would have been without the project (Savedoff *et al.*, 2006). Why have these two movements failed, so far, to reinforce each other?

I argue that this is not an accident. Features of the Goal Movement – as codified in the Millennium Development Goals (MDGs) – have partially obstructed the Evaluation Movement. I suggest that this arises from the incentives faced by proponents of the two movements, and I illustrate the conflicts between the two movements with a case study of how impact evaluation is done in one major aid project now underway in Africa. But things can get better. I will propose ways in which both the Goal Movement and the Evaluation Movement can change to become more complementary.

8.2. The two movements: Goals and Evaluation

The Goal Movement that swept through the development policy world in the late 1990s is the latest manifestation of a recurring pattern: when political change threatens aid budgets, aid agencies justify their spending by linking it to measurable development outcomes.

An early example was the Pearson Commission report of 1969, *Partners in Development* (described in Clemens and Moss, 2007). The Pearson Commission was formed by World Bank President George Woods in 1967 as an explicit response to "concern about flagging enthusiasm among rich countries for making resources available for international development" (World Bank Group Archives, 2003). The report set quantified and time-bound development goals to be met by 1980, including an economic growth target for developing countries. When the next World Bank President Robert McNamara spoke shortly after at Columbia University (World Bank, 1970), he described the "impact" of the Pearson Commission by listing donors' new aid commitments totalling several billion dollars. He argued that alongside the growth target, donors should pledge several measurable declines in poverty indicators such as malnutrition. The most important missing ingredient, he said, was "the dedication, the drive, the determination to see the task through".

A similar process happened in the late 1990s. The OECD (1996) issued a report, *Shaping the 21st Century*, in response to sagging public support for overseas assistance at the end of the Cold War. As a political tool to "sustain and increase the volume of development assistance", the report proposed a small set of quantified, time-bound development targets. These goals, with some additions and changes, were later endorsed by the largest gathering of heads of state in modern history – at the United Nations in 2000 – and named the Millennium Development Goals. They include the halving of poverty and the achievement of universal primary school completion by 2015. The Millennium Declaration, adopting the goals as official UN targets, frames the achievement of the goals – as McNamara did 30 years before - as a function of the "resolve" and "determination" of governments. There is substantial evidence that the Goals succeeded in raising aid budgets and channelling more aid to countries further from the targets (Kenny and Sumner, 2011).

The recent ascent of the Evaluation Movement began around the same time as the most recent wave of the Goal Movement, in the mid-1990s, among a group of academic development economists based primarily in the United States (e.g. Duflo and Kremer, 2005; Gertler *et al.*, 2011). They began to measure the effects of

development projects with more reliable analytical tools adapted from the fields of psychology, public health and labour economics. These studies often gave radically different results from the more traditional qualitative, retrospective, or anecdotal evaluation methods (e.g. Glewwe *et al*., 2004; Banerjee *et al*., 2010). This has contributed to a parallel movement in the development policy world to justify antipoverty spending by focusing on projects that have greater, well-measured impacts for scarce resources (Savedoff *et al*., 2006; Székely, 2011).

The Goal Movement and the Evaluation Movement have much in common. Both have helped shift aid practitioners' attention from project process (such as schools built) toward development outcomes (such as child learning). Leaders of both movements would like to see a large share of aid interventions rebuilt around achieving measurable results (e.g. Sachs, 2006; Banerjee, 2007). But below the surface lie important conflicts between the Goal Movement in its recent manifestation and the Evaluation Movement. Features of the Millennium Development Goals present a powerful obstacle to wider use of rigorous impact evaluation, and thus prevent the two movements from reinforcing one another. To explain how this happens, and how things might get better, we need an understanding of the incentives faced by advocates.

8.3. How conflicts between goals and evaluation can arise

Potential conflicts between development goals and impact evaluation are described in the model of advocacy and evaluation due to Pritchett (2002). He begins with the plausible assumptions that: 1) rigorous evaluation requires the cooperation of project advocates; 2) advocates care more about development outcomes than the public that has to fund them; and 3) advocates' dual objective is to improve development outcomes *and* raise money. Advocates can raise money either by "evaluation" – rigorously demonstrating that their impacts exceed those from alternative aid projects – or by "persuasion" – winning over funders with exaggerated claims of impact.

The model predicts that the amount of learning through rigorous impact evaluation will be less than socially optimal: for advocates, it can "pay to be ignorant". Projects with little impact can prevent careful evaluation from occurring and face little incentive to spend scarce resources on evaluation. But beyond that, even projects with positive impacts – even if evaluation were *free* – might prefer persuasion over evaluation if they face funders who place a low weight on development progress. Such funders might include swing-voter taxpayers in donor democracies. Duflo and Kremer (2005) point out an additional political problem: if enough advocates choose persuasion over evaluation, the field can enter a low-level equilibrium: funders that prefer advocates seeking maximum impact divert support to advocates who persuade *via* exaggeration, which further lowers the returns to evaluation for any individual advocate.

The collision of the Goal Movement and the Evaluation Movement can usefully be seen through the lens of this model. When the Cold War ended in the early 1990s, there was a decline in the importance of overseas development efforts to donor-country taxpayers. Those still willing to fund aid demanded more evidence of results. Advocates who cared more about those results than the average taxpayer, and wished to defend aid budgets, could respond in one of two ways: evaluate or persuade. The more socially beneficial option would be rigorous evaluation – identifying projects with low impact, diverting funding towards those with higher impact, and thereby raising taxpayers' willingness to pay. This socially beneficial approach would be popular among academics whose funding depends more on the rigour of their approach than on the outcome of the evaluation.

But this socially beneficial option would represent a pure, private cost to numerous development advocates. They would not benefit from the information generated by rigorous evaluation if they were already convinced of the effectiveness of their project, and they would only stand to lose

funding, unless their true impacts were sufficiently high to catch the attention of median-voter taxpayers who place relatively little importance on overseas development outcomes. Both of the advocates' dual objectives would thus be harmed by greater use of rigorous evaluation. As funders become increasingly sceptical, most advocates would choose persuasion over evaluation. Again, as Pritchett points out, this could be optimal even for advocates whose projects' true impact is positive, if funders do not care to fund moderately effective projects – either because funders are insufficiently altruistic or because funders prefer to support other projects willing to persuade with exaggerated claims.

An alternative way for advocates to respond to the political changes of the 1990s would be to formulate a set of development goals. In Pritchett's model, the role of these goals would be to convince increasingly sceptical taxpayers that their money was being used to achieve measurable changes in development outcomes. But the model makes a critical prediction about how this would occur: for numerous advocates that would benefit more from persuasion than evaluation, such goals would best be designed not only to persuade without rigorous impact evaluation, but to persuade in ways that prevent rigorous impact evaluation.

8.4. Goals that undermine good impact evaluation

This model can explain key features of the Goal Movement and the Evaluation Movement as they exist today. It can explain why, this far into the results-oriented Goal Movement, so few projects receive rigorous evaluation: advocates can control whether that evaluation occurs, and encouraging rigorous evaluation is privately optimal for few of them. The model also explains why the Goal Movement has taken a form that creates obstacles to rigorous impact evaluation.

The Millennium Development Goals fit this description. They were designed to raise funds for aid projects by highlighting results and asserting that knowledge and institutions exist to achieve those results. They have successfully done so. But they have at least three traits that impede rigorous impact evaluation.

First, the goals treat changes in outcomes as technological problems devoid of social context. The targets in the Millennium Development Goals are the same for all countries. All countries pledge to achieve 100% primary schooling completion by 2015, whether they are starting at 30% or starting at 90% – even if meeting the goal would require faster increases in schooling than any on record (Clemens, 2004).[58] This necessarily frames achieving the goal as a technical problem, to be solved with sufficient effort, expenditure and dedication. Efforts to measure the "cost" of the MDGs embody this assumption. For example, the "cost" of achieving universal primary schooling is measured as the technical cost of paying for the inputs required to teach each child, *assuming that the social, political, and economic problem of getting them into school has been solved* (Clemens *et al.*, 2007). No effort to "cost" the MDGs has attempted to measure the quantity of money that, if spent, would solve the social, political and economic problem of getting children into school. That number, unlike the technical cost of schooling once it occurs, is unknown.

This framing is deadly to proper evaluation of the impacts of efforts to meet the goals. To see why, we need a critical distinction between two types of effects in the impact evaluation literature: the difference between *technical efficacy* (called the "Treatment-On-Treated" or TOT effect) and *project effect* (called the "Intent-To-Treat" or ITT effect).

Suppose an aid project seeks to treat 100 children with a pill. The TOT effect is the effect of the pill on the average child *to whom it is successfully and properly*

[58] Although many of the MDGs are strictly global goals as originally formulated, they have been almost universally interpreted as country-level and even sometimes village-level goals by the United Nations, the World Bank, the Millennium Project, the World Health Organization and other aid agencies.

administered, which may or may not include all 100 children. The ITT effect is the average effect on all children in the original group of 100, *including children who for any reason did not receive the pill or were not properly administered the pill*. For example, suppose a pill cures an infection in 50% of children who properly receive it in ideal conditions, but in the field, only 40% of the children that a project seeks to treat with the pill do end up getting properly treated – due to limitations in demand by parents, logistics of distribution, poor training of health workers, climatic conditions that degrade the pill's effectiveness, or other reasons. In this case, the TOT effect is to cure 50% of children *treated*. The ITT effect is to cure 0.5 × 0.4 = 20% of children *intended* to be treated. Both of these effects are of interest to an implementer and funder, and both are "the effect" of the intervention, in two different senses. But one is multiple times larger than the other.

Because the goals are framed as technical problems devoid of political and economic context, we should expect evaluations of efforts to achieve the goals to focus on the technical efficacy of treatments that the project hopes to administer, not the effect of the project itself. That is, we should expect to see evaluations of TOT effects, not ITT effects. Relatedly, we should expect to see cost-studies of efforts toward the MDGs to focus on the cost of achieving the TOT – for example, the cost of the pill taken by each child to whom it was successfully and properly administered – not the cost of achieving the ITT – the cost of operating the entire project, including the cost of trying and failing to reach children that were not treated.

Most funders only care about ITT impact. Funders who want improvements in development outcomes only care about the answer to the question: "If my money is spent on this project, what will be the overall effect?" (that is, the ITT effect). But because TOT effect per unit of TOT cost will almost always be much larger than ITT effect per unit of ITT cost, advocates asserting to report "results" of their projects will face a strong incentive to describe their "impacts" in TOT terms and to confuse funders about the large differences between the two. If Advocate A advertises, "at 10 cents per pill, 1,000,000 children can be treated for just USD100,000" (a TOT benefit) and Advocate B advertises "spending USD100,000 will result in successfully treatment of 150,000 children" (an ITT benefit), then Advocate A will attract more funding as long as the fundamental difference between these claims is muddy in funders' minds. Clarifying the distinction would go against the interests of Advocate A. The charity appeals with which I am familiar are usually phrased in the manner of Advocate A, even though only Advocate B is explaining to funders the true *result* of spending their money.

Second, the goals are time-bound to create a sense of urgent crisis. The MDGs are extremely short-term goals, based on the notion that massive change is possible in development processes that for today's developed countries took many centuries. A sense of crisis has obvious benefits for fundraising; a core tenet of modern marketing is to urge immediate consumer action ("Call now!").

But this feature of the goals militates against rigorous impact evaluation in two ways. For

one thing, it erodes political support for proper evaluation, promoting the idea that there is no time for careful evaluation in an emergency. This serves to mobilise moral condemnation of the Evaluation Movement, portraying them as uncaring bean-counters who dither in the face of tragedy.

For another thing, it encourages myopia in any evaluation of efforts to meet the goals. The long-term effects, institutional sustainability and financial sustainability of efforts to meet the goals would be little considered in any evaluation. All such concerns are irrelevant to short-term, time-bound goals. Woolcock (2009) points out that rigorous evaluation must be built around the expected time-path of results from that project, a time-path whose medium- and long-run aspects are invisible to the MDGs. The consequence is to distort impact evaluation:

> As things currently stand, however ... international mandates to achieve 'targets' (such as the Millennium Development Goals) generate a net effect, in which the development industry ends up reverse engineering itself, strongly preferring 'high initial impact' projects over projects that might actually respond to the problems that poor countries themselves deem a priority (p.7)

Third, the goals create accountability for inputs, but not effects. The MDGs, like the Pearson Commission goals before them, make donors subject to international shame if aid increases asserted to cause the goals' achievement are not met. But they create no accountability for any of the measurable, time-bound changes in development outcomes they prescribe. Put differently, donor taxpayers and the governments they elect will be subject to embarrassment if aid does not rise, but no person, project, organisation or government will see its prospects changed in the slightest if any of the outcome goals are not met, by any margin. To the question, "Who will lose his or her job if these goals are not met?" the response of the MDGs is silence.

This lack of accountability fosters a culture of impact evaluation in which little premium is placed on independence and transparency. The easiest way to assess a project's "effects" for funding decisions is to use evaluations performed by employees of the project, using confidential data making moralistic assertions about the imperative to carry out one of many antipoverty solutions, based on first-hand anecdotes and opaque statistics. Deaton (2008) summarises this approach as "[t]echnical solutions buttressed by moral certainty" (p.1536). If project leaders face no accountability to deliver results, there is little reason for them to support independent assessments of impact based on transparent outcome data and careful analysis.

The absence of accountability in the MDGs promotes this result. If there is no cost to project advocates from assessments of project impact that differ from their own, there is no incentive to take the trouble to set up independent evaluation mechanisms. Corroborating this idea, among the slim fraction of aid projects that are rigorously evaluated, only a miniscule number are evaluated using independently collected data and/or independent analysis (Savedoff *et al.*, 2006).

8.5. A case study in the collision of goals and evaluation

This story becomes clearer if we consider a specific case where the Goal Movement and the Evaluation Movement have collided. The case suggests how particular traits of the MDGs can impede rigorous impact evaluation. It also illustrates how the two movements could become more complementary.

The Millennium Villages Project (MVP) is a package of simultaneous development interventions in isolated rural village clusters at 14 sites across Africa. The package contains interventions in five areas: agriculture, infrastructure, health, education and water/sanitation. It interprets the MDGs as village-level goals, and seeks to "meet the Millennium Development Goals" at each village site. It is built around the idea of synergies between these different components, and promises to demonstrate that "with adequate resources, people in the poorest regions of rural Africa can lift themselves out of poverty in five years' time" by sparking "self-sustaining economic growth" in remote rural areas, lastingly freeing them from "poverty traps" (MVP, 2011*a*: 1). The project is based at Columbia University and its evaluation is carried out internally. The United Nations is an implementing partner in the project, and both the current and previous Secretaries General have personally and prominently endorsed it. The MVP is the flagship UN antipoverty initiative to emerge from the Millennium Summit.

The first intervention began in Sauri, Kenya, in July of 2004, and subsequently spread to sites in nine other countries. In June of 2010, the project released its first report on the impacts of the intervention (MVP, 2010). This report, like other publications of this project based at one of the world's top institutions of academic research, stressed the project's dedication to rigorous evidence and scientific evaluation.

The 2010 report grossly exaggerated the true impact of the project on the village clusters under intervention (Clemens and Demombynes, 2011). It did this by attributing to its own impacts the entirety of changes, in several instances, that were occurring across the countries and regions where the project is working – and would therefore very likely have occurred had the project not existed.

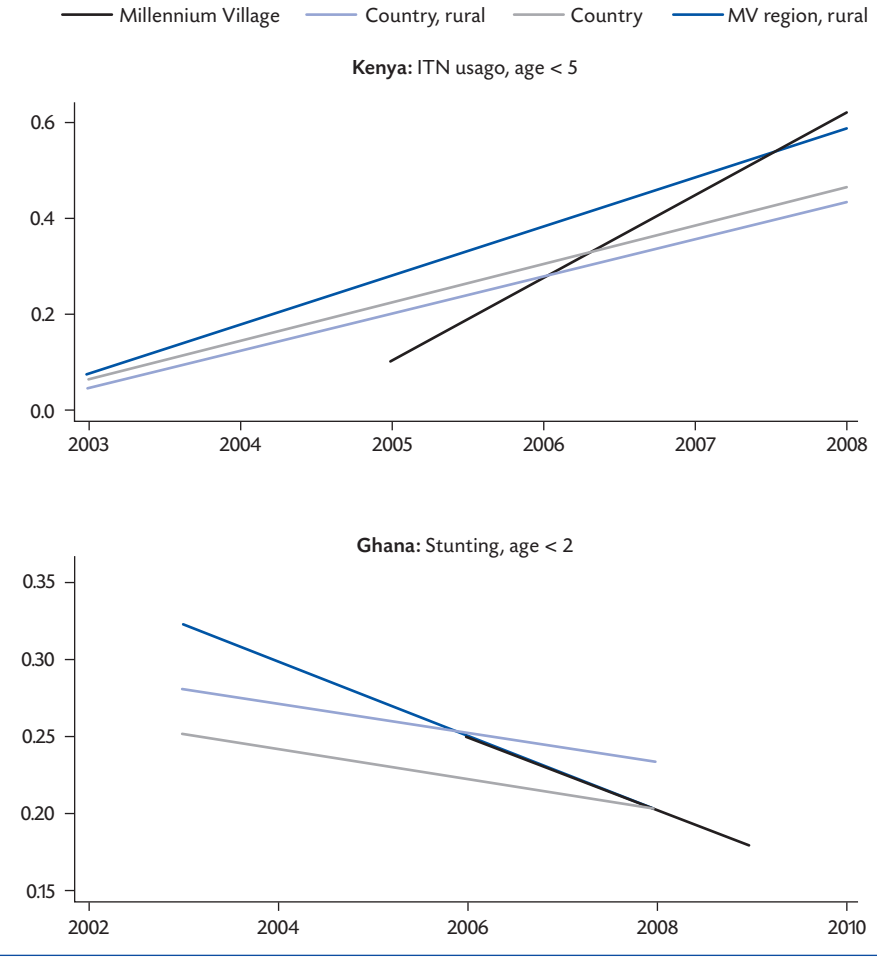
For example, it described as an "impact" of the project the full 52 percentage-point rise in child usage of insecticide-treated bednets in Sauri, Kenya, while child bednet usage rose at similar rates over the same years in the entire surrounding province of Nyanza and the nation of Kenya as a whole.[59] Another

[59] Sauri constitutes less than 2% of Nyanza's population.

"impact" of the project was given as the full 7 percentage-point decrease in chronic child malnutrition (stunting) in Bonsaaso, Ghana, while child malnutrition fell at similar rates across the surrounding region of Ashanti and the entire nation of Ghana.[60] Figure 13 illustrates these how these changes, claimed in full as impacts of the project, in fact represent much broader trends. The report goes on to describe thirty-eight other

Figure 13

Changes at Millennium Village sites, described as "impacts" of the project, compared to changes in areas untouched by the project



Source: Clemens and Demombynes (2011).

[60] Bonsaaso constitutes roughly 1% of Ashanti's population.

changes seen at the project sites as "impacts" of the project. These include – perhaps most inexplicably – increases in cellular phone ownership, in countries that were undergoing explosive nationwide increases in cellular phone ownership during the same period.

To date the project has not retracted or modified any of the precise, quantitative claims of impact made in the original report (MVP, 2010). It ignored the analysis in Clemens and Demombynes (2011) and has continued to claim that the full, quantitative changes in numerous development indicators at the sites represent the "result" and "achievement" of the project (MVP, 2011*b*).[61] Based on those incorrect claims, the project has attracted over USD72 million from the Soros Foundations and recently USD18 million from the UK government, as well as support from the Japanese government and corporate partnerships with Nestle, Panasonic and other multinational corporations.

What the project did instead was to issue a series of public statements that illustrate how the core tenets of the MDGs obstruct a more rigorous evaluation of the project's impacts. As before, I will divide these obstacles to impact evaluation into three types: those that arise from: 1) the lack of social and political context in the MDGs; 2) the time-bound urgency of the MDGs; and 3) the lack of accountability for results in the MDGs.

8.5.1. Obstacles arising from the lack of social and political context in the MDGs

I argued above that the lack of context in the MDGs defines meeting the goals as a technical problem, and would lead evaluation efforts to focus on technical efficacy (TOT effect) rather than the project effect (ITT effect). Advocates would face a strong incentive to confuse the two effects in fundraising efforts, even though this might result in a large exaggeration of the project effect.

The MVP did confuse the TOT effect and the ITT effect in its response to the revelation that many of the "effects" or "achievements" the MVP claimed for itself were occurring in surrounding areas and nationwide. The project objected that trends inside and outside the intervention sites cannot be compared – even in principle – because other interventions are occurring outside the sites. For example, the project leadership wrote:

> These are not pure "control" villages in the sense that governments, NGOs and other development partners are currently involved in scaling up many of the same interventions contained in the Millennium Villages package. . . .untouched comparison communities, if they ever existed, are unlikely to exist now. In real-world settings pure "controlled conditions" for well-established interventions are impractical.[62]

[61] To claim a result is synonymous with claiming an effect or impact. It is nonsense to claim, as a "result" of a project, something that occurred alongside the project but was not caused by it. For the same reason it is nonsense to say, "I took my umbrella to work today, and as a result, it rained".

[62] Paul Pronyk, John McArthur, Prabhjot Singh and Jeffrey Sachs, "Evaluating the Millennium Villages: A response to Clemens and Demombynes", Millennium Villages blog, October 13, 2010.

And Jeffrey Sachs and Prabhjot Singh wrote that comparing the intervention sites to others rests on "flawed" logic:

> In a single-intervention study at the individual level (e.g. for a new medicine) one can have true controls (one group gets the medicine, the other gets a placebo or some other medicine). With communities, there are no true controls. Life changes everywhere, in the MVs and outside of them.[63]

These observations reflect a failure to grasp the difference between the TOT effect and the ITT effect. A "pure" control group – villages with no interventions of any kind, by anyone – would only be necessary if evaluators sought to measure the technical efficacy of the different components of the successful application of those components (the TOT effect). But the evaluation of a *project's* effect is an ITT effect. The counterfactual for a project is, by definition, everything that would have occurred in the absence of the project. That necessarily includes all technical interventions by any other actor that would have occurred anyway.[64] As evaluation expert David McKenzie observed, the fact that "life changes everywhere" is precisely the reason why it is essential to compare intervention sites with other sites.[65]

In a further sign of the evaluation confusion encouraged by the MDGs, the project responded that impact evaluation was unnecessary, because rigorous impact evaluation is only informative when an intervention is "unproven", and many elements of the package intervention are ostensibly "proven":

> [P]rogress towards the MDG targets is less about designing novel interventions and technologies, and much more about creating effective local systems to put these proven interventions into practice. The main research questions are not simply "does it work?", but rather how to overcome complex implementation and financial challenges in a diverse range of poor and hard-to-reach communities.[66]

Again, this reflects a fundamental confusion reflecting the MDGs' *definition* of development progress as a technical problem. Asking the question "Does it

[63] Jeffrey Sachs and Prabhjot Singh, "Learning in and from the Millennium Villages: A response to Lawrence Haddad", Millennium Villages Blog, October 16, 2011. (At the time of writing this post has been suppressed and deleted from the Internet by its authors, but most of its text is preserved within a critique by David McKenzie at http://blogs.worldbank.org/impactevaluations/jeff-sachs-the-millennium-villages-project-and-misconceptions-about-impact-evaluation.)

[64] In measuring the ITT effect, we should only be concerned about interventions in comparison areas if there is a clear indication that those interventions would not have occurred at the project intervention sites if the project did not exist. In the MVP case, there was no such indication. For example, the project claimed the full increase in school enrolment at Bonsaaso, Ghana, as its own "impact" (MVP, 2010). But there is little reason to believe that Ghana's nationwide campaign to raise school enrolments during the same period – starting with primary school fee elimination in late 2004 – would have affected Bonsaaso differently from other communities had the MVP never occurred.

[65] David McKenzie, "Jeff Sachs, the Millennium Villages Project, and Misconceptions about Impact Evaluation", World Bank Development Impact Blog, October 19, 2011.

[66] Pronyk, McArthur, Singh and Sachs (2010), *op. cit.*

work?" about any technical element of the project is a TOT effect. It is a fundamentally different question to ask whether the entire project, including any perfect or imperfect application of all of its technical components achieves its stated goals of changes in outcome. This is an ITT effect. The project effect (ITT effect) can remain entirely unknown even if the technical efficacy (TOT effect) of every one of its components is known with certainty.

The MVP's (2011*a*) claims to lift remote rural African communities out of poverty in five years time, by sparking self-sustaining economic growth and freeing villages from poverty traps is not a technical claim about fertiliser, bednets, or any other component of the project. It is not a claim of TOT effect. It is a claim of ITT effect. The claims of overall impacts of the project in its evaluation reports are claims about ITT effect. To defend a project's ITT effect based on assertions about its TOT effect is a basic mistake and misleading to funders.

The project raised a related objection to comparing outcome trends at the sites to trends elsewhere: remarkably, it claimed responsibility for many of the changes occurring *nationwide* in the ten countries where it works. To defend against a critique of before-and-after impact evaluation by *The Economist*, which compared trends at the intervention sites to national trends,[67] Jeffrey Sachs wrote: "The project itself has been encouraging the take-up of a range of interventions (bed nets, fertilizer, high-yield seeds, new diagnostic methods, and so forth) in neighbouring villages and at the national scale".[68] If it were true that this project caused a large portion of the region-wide and nation-wide changes shown in Figure 13, then it would indeed be incorrect to compare an intervention site to the surrounding region and nation; trends in both areas would capture the effect of the project.

But this would mean that that some large fraction of development improvements at the national level would not have occurred if the Millennium Villages Project had not occurred. Such a claim requires strong evidence for a project that asserts its claims of impact reflect "peer-reviewed science".[69] If the project had never occurred, would there have been large changes in *national-level* trends in schooling, malnutrition, malaria, vaccination, skilled birth attendance, and the many other outcomes whose changes at the project sites have been claimed in full as impacts of the project? Many of the positive trends occurring nationwide in the countries in question, such as Ghana's universal schooling campaign or Kenya's anti-malaria campaign, began many years before the MVP existed. And the Millennium Villages are tiny areas of intervention: for example, the project intervention site at Bonsaaso contains 0.1% of the population of Ghana.[70] It is unclear

[67] The Economist, "The big push back: Randomised trials could help show whether aid works", December 3, 2011.

[68] Jeffrey Sachs, "Challenges at the Cutting Edge of Fighting Global Poverty", *Huffington Post*, December 4, 2011.

[69] Jeffrey Sachs, "The Millennium Villages Project is working well", *The Guardian* Poverty Matters blog, October 13, 2011.

[70] http://mp.convio.net/site/PageServer?pagename=mv_bonsaaso

by what mechanism the project's efforts there could have markedly accelerated national trends in any of the development outcomes under consideration. The burden of proof certainly lies on anyone outside Africa who would claim individual responsibility for substantial portions of the revolutionary improvements in living standards that have accompanied a continent-wide economic renaissance over the last decade (e.g. Radelet, 2010).

A further consequence of the MDG framing of aid interventions as technical, acontextual solutions is a lack of concern about the external validity of evaluation results. The Millennium Village intervention sites were chosen specifically because the designers thought that the project might work better at those sites than at other sites. It is therefore difficult to infer from changes at the demonstration sites how the same project might affect other areas. When we pointed this out, the project protested:

> Clemons [*sic*] and Demombynes claim that the choice of villages was somehow "subjective" rather than rigorous and evidence-based. In fact, this issue has already been discussed at length in a peer-reviewed and registered evaluation process (The Lancet, protocol number 09PRT-8648). ... Sites were chosen "purposively" to represent over 95% of the agro-ecological zones on the continent - reflecting a variety of systems-level challenges, disease profiles, and baseline levels of infrastructure and capacity. Within each country, selection criteria included rural areas with high rates of poverty and where at least 20% of children were undernourished. ... These data refute any insinuation that the Millennium Villages were somehow systematically advantaged at the outset of the project.[71]

This confident statement is directly contradicted by the same research protocol it cites. That protocol reads: "The non-random selection of intervention communities has the potential to introduce bias ... Issues of feasibility, political buy-in, community ownership and ethics also featured prominently in village selection for participation".[72]

The existence of competent and cooperative local partners is key to any project's success, and those who have worked in the field know that it can be rare for the right elements to come together in any given community. The project's own documentation then gives good reason to believe that its effects might be less in communities that were not specially selected to have competent and cooperative local partners. But when the problem of meeting the MDGs is framed as a technical problem to be solved *after* all political and social obstacles to implementation have been solved – as the MDGs do – technical evaluation should be unconcerned by such limitations, and results from one social, political and economic setting should be

[71] Pronyk, McArthur, Singh, and Sachs 2010, *op. cit.*

[72] Columbia University, "The Millennium Villages Project: Assessing the Impact on Child Survival and the Millennium Development Goals in Sub-Saharan Africa (MVP)". For the full protocol follow the links at http://clinicaltrials.gov/ct2/show/NCT01125618

more readily assumed to apply wholesale to a different setting.

Finally, I discussed above how the MDGs distort the discussion of project costs. Costs are an essential component of useful impact evaluation, since there are competing potential uses for every aid dollar. Costs go unmentioned in the MDGs, except to the extent that aid must ostensibly rise by some amount in order to achieve the goals. If improving development outcomes is a technical problem of implementing interventions for which the TOT effect is known – as the MDGs define the problem – cost is only relevant for donors to the extent that it indicates whether they can or cannot afford to pay for the technical interventions that achieve the goals. MVP publications universally discuss cost in this light, portraying cost as low in some unspecified, absolute sense.

But the TOT effect per unit cost is not helpful to funders choosing between projects, as they should, according to the ITT effect per unit cost. The MVP does not undertake any analysis of the effects of alternative uses of the same money spent on the project. The cost of the MVP intervention is high: on average, the intervention requires on-site expenditures of USD160 per year, for every man, woman and child at the intervention sites.[73] While the project's public documents are not clear about what this number includes, it does not appear to include many off-site costs, such as the office space of the Earth Institute in New York that is devoted to the project.

The project has not publicly released any analysis, at the time of this writing, on the cost-effectiveness of its antipoverty intervention. This makes its economic impacts opaque. But even if the impacts claimed by the project were rigorously assessed, they would be uninformative about whether the project is efficient or wasteful in achieving those impacts.

The first thing to note about the cost of the project is that this on-site expenditure represents a gigantic intervention in the local economy. For example, income per capita at the project's flagship site, Sauri, Kenya, is roughly USD145 per year.[74] In other words, the annual MVP intervention is enormous in context. It is larger in economic terms than the entire local economy.

Very large effects can and should be expected from interventions of this magnitude. If the same money were simply handed out as cash for the duration of the intervention, it would more than double local income per capita, with numerous consequent improvements in education, health and other social indicators that typically arise from unconditional cash transfers to the very poor (e.g. Baird *et al*, 2011). In order for the project to be a superior intervention against income poverty relative to distributing cash, it must raise incomes either to a degree or duration (or

[73] MVP, "Sustainability and Cost", MVP website accessed January 17, 2012 at http://millenniumvillages.org/the-villages/key-activities/sustainability-cost

[74] The latest estimate of per-capita income in Siaya district is KSh10,784 in 2005 (http://mirror.undp.org/kenya/UNDP_4thKHDR.pdf), a time when the exchange rate was about KSh75=US$1.

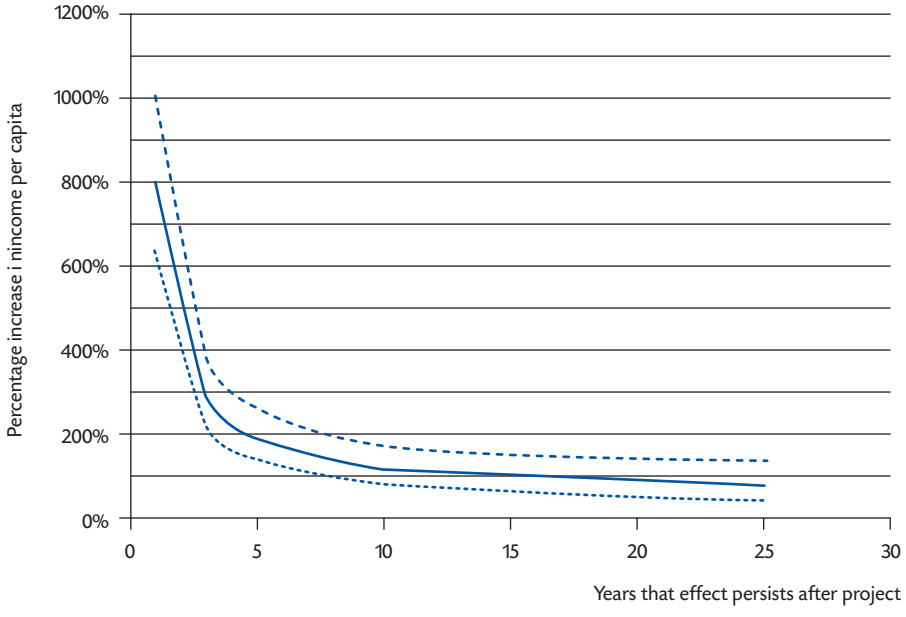
both) that makes the present value of the economic effect exceed the cost.

Figure 14 shows the results of a simulation of the conditions under which the project is a superior antipoverty measure to direct distribution of cash. It conservatively assumes that local income per capita starts at USD150, and the project costs USD150 per year. The vertical axis shows the percentage increase in local incomes caused by the intervention. The horizontal axis shows the number of years after the conclusion of the five-year intervention that the effect is sustained. The line traces out the set of points where the project "breaks even" – that is, where the present value of the cost of the project (5 years times USD150/year) equals the present value of the increased income stream brought about by the intervention. The three different lines reflect different assumptions about the interest rate used to discount future income. If a point is above or to the right of the line, the project creates more income than it costs. If a point is below or to the left of the line, the project destroys value; it costs more than one dollar to create one dollar of added income for impoverished families.

The simple analysis in Figure 14 suggests that, if the project raises local incomes by anything less than 100%, that effect must persist for at least ten years after the project

Figure 14

Break-even simulation for costs and economic benefits of project



Source : author

stops pumping resources into the local economy. If the ten-year sustained increase is any smaller, or if the 100% increase is sustained for any less than ten years, the local population would have experienced a greater economic benefit from a simple cash transfer in the amount of the project's cost. Put differently, this would mean that switching from a cash transfer to the MVP would economically harm local residents while it economically benefits people employed by the project.

This back-of-the-envelope calculation obviously requires refinement. But it does suggest that the cost of the project is large enough to raise serious doubt about whether it is better than cash transfers – in the absence of giant increases in income lasting for a decade or more after the intervention ends. At this point there is no evidence of any kind that such massive and lasting increases in income will occur at the sites.

Because (at the time of writing) the project has not released any data regarding the effects of the project on income, a transparent assessment of the economic costs and benefits of the project is not possible. Because the project promised to lift people out of poverty in five years' time, and began almost eight years ago (July 2004, in Sauri), it is notable that the project's evaluation reports to date have not contained any data about income trends at the sites (MVP 2010, 2011*b*), despite the fact that the project does collect income data, and has selectively reported numerous other non-income indicators that it collects. One evaluation study, conducted independently of the MVP and without its knowledge or cooperation, compared treated households in Sauri to nearby and otherwise similar untreated households, finding no significant difference in income (Wanjala and Muradian, 2011). Clearly more evidence is required. The burden of proof lies on the project to demonstrate what the returns per dollar of the project have been.

8.5.2. Obstacles arising from the time-bound urgency of the MDGs

The Millennium Declaration and the MDGs portray the problem of underdevelopment as an urgent crisis that, with sufficient "determination", can be solved in a few years. This framing of the problem serves to erode support for careful evaluation of any related projects, as one would ridicule a painstakingly precise evaluation of emergency food aid in an acute famine.

On these grounds, the MVP has categorically rejected the idea that its claims of long-term impact undergo any evaluation at all before massive new resources are pulled away from alternative projects to be devoted to the MVP. The MVP (2011*a*, 2011*b*) has stated that its effects will be self-sustaining and will outlast the project. Because other village-level package interventions have seen their effects quickly disappear after external money stops coming in (Chen *et al*., 2009), Clemens and Demombynes (2011) recommend that impact evaluation test these claims of sustained impact at a point 5–10 years after project completion – that is, 10–15 years after the project began – before the project is massively expanded at the expense of alternative uses of aid money. The project mocked this idea:

> Clemens and Demombynes also suggest that efforts to take interventions to scale should wait at least 15 years until evidence of long-term effects can be proven and sustained. This assertion cannot be taken seriously. ... It would be the height of folly to delay ... Economists like Clemens and Demombynes should stop believing that the alleviation of suffering needs to wait for their controlled cluster randomized trials.[75]

While it is clear from the above that the authors consider anyone requiring evidence of their strong claims to be immoral promoters of suffering, it does not answer the key question: if a project claims to have impacts that last long after its intensive five-year intervention, how is it possible to evaluate those impacts in less than 5–10 years after the project ends?

The MVP goes out of its way to state that it is not "charity"; it is a project to create *lasting* freedom from poverty traps, not an emergency humanitarian project to end suffering. The effects of this project after the intervention ends are not known and cannot be known without evaluation years after it ends. The project was not described to its funders as a project to end suffering temporarily while large amounts of money were flowing; suffering that would return immediately after it ended. Rather, it was described as a project that, with a one-time intensive intervention, would cause lasting change. Such a claim can only be assessed with an impact evaluation on the timescale of the stated goals.

But the MDGs, in order to urge action by potential funders, redefine the problem of development as a short-term problem. If "development" and "poverty alleviation" are reduced to meeting very short-term targets on specific indicators, the only relevant form of impact evaluation is a form that assesses short-term impacts. Myopia of this sort is a reasonable response to the incentives created by the MDGs, which say nothing at all about long-term development trajectories, how those might be changed, or how such changes might be assessed.

Another symptom of the MDGs' short-term, moralistic framing of the development problem in the MVP evaluation is the project's pattern of resistance to comparing treated and untreated villages – a stance that makes sense in the context of a crisis or emergency. Three years into the project, in 2007, its leaders maintained the policy that no data would be collected at sites not receiving the intervention. They believed that such data collection was unethical, writing:

> For ethical and practical reasons, there are no formal 'control' villages. Instead, project impact is assessed by rigorous before-and-after comparisons and detailed studies by sector. ... The ethical reasons relate to the fact that many core interventions (e.g., malaria control, access to safe water) are life-saving and would be ethically inappropriate to deny in a control village. (Sanchez *et al.*, 2007, p.16779)

[75] Pronyk, McArthur, Singh, and Sachs, 2010, *op. cit.*

This does not follow. Collecting data from an untreated area is not the same thing as "denying" the intervention – except in an emergency with unlimited resources, where the only obstacle to treatment is "determination". Unless there is sufficient funding and other resources to offer the intervention to the entire population simultaneously, the existence of untreated areas is a fact that the project cannot change. An outcome that is beyond the project's control cannot be described as the responsibility of the project. Certainly the MVP had nothing close to the resources necessary to treat even one entire province of one country, much less the entire populations of ten countries. Gathering information on outcomes beyond the project's control does not implicate the project any more than a war correspondent's work makes him or her responsible for warfare.

The project later revised this stance and, in 2008, issued a research protocol requiring untreated comparison sites.[76] But there was lasting damage to the impact evaluation, because the project's refusal of comparison sites meant that no "baseline" (initial) data were collected at many of those sites. This has prevented comparison of trends at treated sites to trends at untreated sites during the same time period. This is another example of the direct harm to careful evaluation done by the short-term-crisis mentality underlying the MDGs.

Another common objection, also voiced by the MVP, is that rigorous impact evaluation is extraordinarily costly. Rigorous impact evaluation is simply impact evaluation that considers seriously and objectively what would have happened without the project. One easy way to do this, as we did in the graphs of Figure 13 and others like it, is to compare trends at the intervention sites to trends away from those sites. All of the data we used to conduct that comparison were available for free online, as is the software necessary to conduct analysis. In many situations, careful consideration of the counterfactual can cost little or nothing. Random selection of the intervention sites, which would have greatly raised the credibility of the MVP impact evaluation results, is also costless in and of itself. Intervention sites must be chosen by some method, and selecting them at random is a simple matter of having a spreadsheet generate random numbers.

8.5.3. Obstacles arising from the lack of accountability in the MDGs

The MDGs are silent on the subject of who will be held responsible if any of the changes in development *outcomes* – poverty, schooling, mortality – are not met. Because no person or organisation therefore has an extrinsic stake in the outcome, this serves to undermine support for independent, objective, transparent assessments of outcome that are the *sine qua non* of rigorous impact evaluation.

Again, the stance of the MVP on impact evaluation fits these incentives. The first response of the project to the critiques by Clemens and Demombynes (2011) was to attack the analysis on *ad hominem* rather

[76] Available at http://clinicaltrials.gov/ct2/show/NCT01125618

than substantive grounds. It dismissed what it called "second-hand" or "armchair criticism" on the apparent grounds that impact evaluation using secondary data is not credible regardless of evaluation method.[77]

This assertion is untrue. To take an example from medicine, an evaluation of a chemotherapy drug that proceeds from systematic secondary data on representative individual outcomes with and without the drug can be informative about the drug's effects. Such systematic, secondary analysis would be a more reliable guide to the effects of the drug than a week spent observing first-hand and anecdotally the administration of the drug to a few patients. Careful methods and representative data are much more important to sound conclusions than whether the data used were gathered by the analysts themselves or by someone else. But the MDGs create no incentive to establish accountability for any organisation's success or failure in its efforts – accountability that, like all analysis of new medical treatments' effects by the United States Food and Drug Administration, can and should be second-hand.

But the MDGs are silent on accountability, creating no demand at all for independent analysis. When analysis is not independent, there is the risk of selective reporting of results by analysts under social and economic pressure, even when those analysts are fully honest individuals. This, I believe, does not reflect on the ethics of the researchers, but on fundamental cognitive processes apart from ethics. (I, for example, do not consider myself cognitively capable of writing a fully objective analysis of my own employer, regardless of my inherent honesty.) Some of the "peer-reviewed science" on which the MVP bases its claims of impact shows clear signs of selective reporting of results.

In one such study (Remans *et al.*, 2011), Jeffrey Sachs and several colleagues attempt to measure the effect of the MVP intervention on child malnutrition. They do so by comparing trends in one malnutrition indicator, child stunting (height for age, «2 years), at several intervention sites *during the project* to the national-level trend in child stunting *during the two decades prior to the project*.[78]

As we point out in a comment published shortly afterwards by the same journal (Clemens and Demombynes, 2012), this research method is badly misleading because it involves the selective reporting of results flattering to the project. *Prior to the project*, national trends in child malnutrition were indeed flat for the countries in question for many years. But *during the project*, almost all of these countries were experiencing large national declines in child malnutrition as sub-Saharan Africa entered a period of relative prosperity. Furthermore, the paper tests the effects of the intervention on three different measures of child malnutrition (stunting,

[77] These comments appear in: Jeffrey Sachs, "'Millennium Villages, on Track to Reach 2015 Goals'–Press Conference", at United Nations headquarters in New York (16:30 mark); and Jeffrey Sachs and Prabhjot Singh, *op. cit.*

[78] The data at the project sites represent the period 2005–2008 or 2006–2009, depending on the sites. The data for the national-level trend are intended to cover 1986–2008 (incorrectly labelled in Figure 3 of the paper as covering 1988–2008). But of the 37 data points that establish this trend, 35 come from the period 1986–2006. That is, the national-level trend estimated in the paper contains almost no information about national-level trends during the intervention.

underweight and wasting), finds a significant effect on only one of these (stunting), and reports *only that finding* in its abstract, introduction and conclusion. It reaches its conclusion on stunting based on the incorrect use of levels of statistical significance that would only be appropriate if the effect on stunting was tested in isolation. The project's analysis, carried out internally with data accessible only to the project, therefore tends to substantially exaggerate the impact of the intervention.

Why would the project compare trends at the sites in one time period to trends nationwide in a different and irrelevant time period? We can eliminate two possible reasons. In our own analysis we used data on the same malnutrition outcome, publicly available for many of the same countries during the period of the intervention, and we showed that there were substantial declines regionwide and nationwide for the countries and time period in question. We provided our analysis to three co-authors of the malnutrition study a year before their work was submitted for publication. It is therefore unlikely that they were unaware that nationwide data were available during the period of the project, or unaware of the regionwide and nationwide declines in child malnutrition occurring at the time of the intervention.

It is difficult to see why independent scientists would have chosen to make the same odd comparison that the project's internal evaluation team made, a comparison that resulted in exaggerated estimates of project impact with the trappings of a reputable scientific journal. The authors of the study had an interest in reporting positive outcomes; one, Jeffrey Sachs, wrote a book prior to the beginning of the project in which he stated the firm conviction that the Millennium Village package intervention could and would cause the "end of poverty" (Sachs, 2005); another of the study's authors, John McArthur, was the chief fundraiser for the project as CEO of Millennium Promise. While I certainly do not believe that either of these people had a personal financial interest in the outcome of the research, nor do I believe either of them did anything unethical, it is nevertheless clear that their public promises and professional positions could hardly leave them completely disinterested in the outcome of the scientific research they were called upon to perform. Independent evaluation would have been more credible. But again, the total silence of the MDGs on accountability for results implies that independent evaluation serves no clear purpose.

In response to the above concerns about independent evaluation, the MVP leadership wrote:

There has been much discussion regarding whether an evaluation that is not 'independent' can be truly rigorous. These perspectives were surprising to me coming from the public health field–where virtually all primary research is conducted by the investigators themselves. The amount of oversight within the MV project is quite striking in fact. We are independently overseen by 11 institutional review boards to whom we report annually.[79]

[79] Paul Pronyk, Jeffrey Sachs, and Prabhjot Singh, "Perspectives on Monitoring and Evaluation in the African Millennium Villages", Millennium Villages blog, October 23, 2011.

This fact is not relevant to the independence of the impact evaluation. Institutional Review Boards at universities exist to ensure that human beings who are the subjects of research are not harmed by that research. They do not independently check each scientific inference made by each researcher on a project. They do not run their own statistical analysis of research projects' data, they do not check to see if the conclusions of each paper produced by a project are justified by the evidence, and they typically comprise small interdisciplinary groups with no substantive experience within the research subfield in question. For example, in the child malnutrition study referenced above, it would have been the job of the Institutional Review Board to review project documents looking for any sign that children were harmed by the intervention; it would *not* have been the job of the Institutional Review Board to check that the research team's conclusions from its data were well-founded.

There is no substitute for independent, disinterested analysis. But project advocates have little reason to encourage independent evaluation – and every reason to fight it – as long as they are not accountable for the success or failure of a project to deliver the changes in development outcomes it promises. By omitting any such accountability mechanisms, the MDG vision of development directly undermines independent and rigorous impact evaluation.

Raising money for development interventions frequently requires project implementers to make promises to donors – promises of large impacts. Competition among grant-seekers can lead to pressures for overpromising larger and larger impacts. This inherently creates pressure on implementers to report the impacts they promised; a set of incentives that can shape research decisions and interpretation. Again, I do not speak of ethical lapses, but rather of a cognitive colouring generated by strong incentives. I believe that even the most ethical scientist could be susceptible to such colouring because it is inherent to human nature. The problem is not moral but institutional, and the solution is institutional: impact evaluation should be executed by analysts independent of the fundraising and implementing apparatus (Savedoff *et al.*, 2006).

A further natural consequence of the MDGs' lack of accountability for specific results is that projects can report success by simply redefining success. The MVP has stated numerous different goals since it began. On one hand, the project states that its purpose is not to prove impacts, because the interventions are already "proven", but rather to "to design and document effective delivery systems" for these "proven" interventions, and that Clemens and Demombynes' (2011) focus on development outcomes "reflects a basic misunderstanding of the MVP's goals and purpose".[80] But these statements are bewildering when the project's evaluation reports (MVP, 2010, 2011*b*) are filled from front to back with quantitatively precise claims that numerous changes in development indicators are the "impact", "result" and "achievement" of the project. If it is true that the goal of the

[80] Pronyk, McArthur, Singh and Sachs (2010), *op. cit.*

project is simply to design systems, the system design is the impact of the project and this is what should be assessed. If it is to affect development outcomes, then this is what should be assessed.

Either one or the other criterion is fine; what is inappropriate is to change the stated goal from moment to moment. What is critical is to choose one criterion of "best" – relative to what – and maintain that criterion. If there are multiple definitions of "best", projects can simply declare that whichever of multiple stated goals was met had been their goal all along. If the goal moves between sparking sustained economic growth, eliminating poverty, emergency reduction of "suffering", generating knowledge of system design, and scores of different development outcome indicators, then proper evaluation and accountability for results is frustrated: if a project fails to meet one stated goal, it can simply stress that the *true* goal was one of the other ones. Independent, disinterested analysts are more likely to avoid sliding from one set of goals to another as a project proceeds, but again, the MDGs define such independence as unnecessary by defining accountability as irrelevant to the grand global project of development.

8.6. Conclusions

There are many ways that the Goal Movement and the Evaluation Movement conflict – as the two movements have evolved so far. This provides one partial explanation for the scarcity of rigorous impact evaluation even at this late stage of the Goal Movement. But this can change, and the two movements could become more complementary.

There is an opportunity, now, to change this dynamic. The Millennium Development Goals expire in 2015, and today there is an active debate about what comes next.[81] There is likely to be a new set of goals, and those goals could be built to help the Goal Movement and the Evaluation Movement complement each other to attract and target funding. This could happen in at least three ways:

Development goals that are *country-specific* would define the development problem to be inherently contextual, rather than a technical problem whose achievement depends solely on dedication and expenditure. Country-specific schooling goals, for example, would be to double past rates of progress or halve the gap between current and universal schooling rates; both of these account for country-specific context. This would draw attention away from evaluations of technical efficacy (Treatment-On-Treated effects) and toward the more policy-relevant evaluations of project effect (Intent-To-Treat effects). It would make cost estimates more informative, shifting evaluation efforts from measuring cost-per-unit-if-successfully-delivered to measuring cost-of-causing-units-to-be-successfully-delivered. The latter is much more useful to funders wishing to meet the new goals and making decisions about how to allocate scarce resources.

Development goals that are *long-term* would lessen the "crisis" case for casting careful analysis aside, and direct more effort toward learning to solve long-term problems. It would assist rigorous evaluation by encouraging evaluations that match the stated goals upon which funders base project allocation decisions, such as deciding between two projects that claim long-term sustained impact. It would reinforce the notion that careful impact evaluation does not stand in the way of ethical advocates taking emergency action; careful impact evaluation is a useful tool for advocates seeking ethically to do the most lasting good with scarce aid resources in places undergoing long-term structural change.

Development goals that specify any degree of *accountability* for changes in outcomes, for any person or organization, would also assist rigorous impact evaluation. When any portion of the extrinsic motivations of advocates hinge directly on results, the stakes are higher and the pressure for

[81] Sumner and Tiwari (2010) and Melamed (2012) survey the latest discussion about what to do next.

objective criteria of success increases. This would create more pressure for independent and transparent analysis of impacts – elements that are critical components of any impact analysis that is to be called rigorous. Accountability, that is to say, encourages learning.

In short, the Goal Movement and the Evaluation Movement need not be antagonistic. They share common objectives: to demonstrate development results to a sceptical audience. The current discussions about new goals present a rare opportunity to bring the two movements into greater partnership.

References

Baird, S., C. McIntosh and B. Özler (2011), "Cash or condition? Evidence from a cash transfer experiment", *Quarterly Journal of Economics,* 126 (4): 1709–1753.

Banerjee, A. (2007), "Inside the Machine: Toward a new development economics" *Boston Review*, 32 (2): 12–18.

Banerjee, A., E. Duflo, R. Glennerster and C. Kinnan (2010), "The miracle of microfinance? Evidence from a randomized evaluation", Working Paper, Massachusetts Institute of Technology, Cambridge, MA.

Chen, S., R. Mu and M. Ravallion (2009), "Are there lasting impacts of aid to poor areas?" *Journal of Public Economics,* 93: 512–528.

Clemens, M.A. (2004), "The long walk to school: Development goals in historical perspective", CGD Working Paper, Center for Global Development, Washington, D.C.

Clemens, M.A. and G. Demombynes (2011), "When Does Rigorous Impact Evaluation Make a Difference? The Case of the Millennium Villages", *Journal of Development Effectiveness*, 3(3): 305–339.

Clemens, M. A. and G. Demombynes (2012), "Multisector intervention to accelerate reductions in child stunting: an independent critique of scientific method", *American Journal of Clinical Nutrition*, March, forthcoming.

Clemens, M.A., C. Kenny and T.J. Moss (2007), "The trouble with the MDGs: Confronting expectations of aid and development success", *World Development,* 35(5): 735-751.

Clemens M.A. and T.J. Moss (2007), "The ghost of 0.7%: Origins and relevance of the international aid target", *International Journal of Development Issues,* 6(1): 3-25.

Deaton, A. (2008), "Maximum Prophet", *The Lancet,* 372 (9649): 1535–1536.

Duflo, E. and M. Kremer (2005), "Use of Randomization in the Evaluation of Development Effectiveness", in G.K. Pitman, O.N. Feinstein and G.K. Ingram (eds), *Evaluating Development Effectiveness*, World Bank Series on Evaluation and Development, Vol. 7, Transaction Publishers, New Brunswick, NJ.

Gertler, P.J., S. Martinez, P. Premand, L.B. Rawlings and C.M.J. Vermeersch (2011), *Impact Evaluation in Practice,* World Bank, Washington, D.C.

Glewwe, P., M. Kremer, S. Moulin and E. Zitzewitz (2004), "Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya", *Journal of Development Economics,* 74(1): 251–268.

Kenny, C. and A. Sumner (2011), "More Money or More Development: What Have the MDGs Achieved?", CGD Working Paper 278, Center for Global Development, Washington, D.C.

Melamed, C. (2012), "After 2015: Contexts, politics and processes for a post-2015 global agreement on development", ODI Research Report. Overseas Development Institute, London.

MVP (2008), *The Millennium Villages: Annual Report, January 1 – December 31, 2008,* The Earth Institute at Columbia University, New York.

MVP (2010), *Harvests of Development in Rural Africa: The Millennium Villages After Three Years,* The Earth Institute at Columbia University and Millennium Promise, New York.

MVP (2011*a*), "Millennium Villages Project: Overview", Millennium Villages Project, New York, http://millenniumvillages.org/files/2011/02/MVInfokit_rev17.pdf, accessed January 17, 2012.

MVP (2011*b*), *The Millennium Villages Project: The Next Five Years 2011-2015,* Millennium Villages Project, New York.

OECD (1996), *Shaping the 21st Century: The Contribution of Development Cooperation,* OECD, Paris.

Pritchett, L. (2002), "It Pays to Be Ignorant: A simple political economy of rigorous program evaluation", *Journal of Policy Reform*, 5(4): 251–269.

Radelet, S. (2010), *Emerging Africa: How 17 Countries Are Leading the Way,* Center for Global Development, Washington, D.C.

Remans, R., P.M. Pronyk, J.C. Fanzo, J. Chen, C.A. Palm et al. (2011), "Multisector intervention to accelerate reductions in child stunting: an observational study from 9 sub-Saharan African countries", *American Journal of Clinical Nutrition* 94(6): 1632–1642.

Sachs, J. (2005), *The End of Poverty: Economic Possibilities for Our Time,* Penguin Press, New York.

Sanchez P., C. Palm, J. Sachs et al. (2007), "The African Millennium Villages." *Proceedings of the National Academy of Sciences*, 104(43): 16775–16780.

Savedoff, W., R. Levine and N. Birdsall (2006), *When Will We Ever Learn? Improving Lives Through Impact Evaluation,* Center for Global Development, Washington, D.C.

Sumner, A. and M. Tiwari (2010), "Global Poverty Reduction to 2015 and Beyond: What has been the Impact of the MDGs and What are the Options for a Post-2015 Global Framework?", Working Paper 348, Institute of Development Studies, Sussex.

Székely, M. (2011), "Toward Results-Based Social Policy Design and Implementation". CGD Working Paper 249, Center for Global Development, Washington, D.C.

Woolcock, M. (2009), "Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy", *Journal of Development Effectiveness*, 1(1): 1–14.

World Bank (1970), Address to the Columbia University Conference on International Economic Development by Robert S. McNamara, President, World Bank Group, February 20, New York.

World Bank Group Archives (2003), "Pages from World Bank History: The Pearson Commission", http://go.worldbank.org/JYCU8GEWA0, accessed February 19, 2012.

Wanjala, B.M. and R. Muradian (2011), "Can Big Push Interventions Take Small-scale Farmers out of Poverty? Insights from the Sauri Millennium Village in Kenya", CIDIN Working Paper 2011-1, CIDIN, Nijmegen.

Authors' Biographies

François Bourguignon

François Bourguignon is the director of the Paris School of Economics. After four years as the chief economist and first Vice President of the World Bank in Washington, he returned to France in 2007, where he took up his former position of professor of economics at the EHESS (School for advanced studies in the social sciences). Trained as a statistician, he obtained a Ph.D. in Economics at the University of Western Ontario, followed by a State Doctorate at the University of Orleans in France. His work is both theoretical and empirical and principally focuses on the distribution and redistribution of revenue in developing and developed countries. He is the author of a great number of books and papers in specialised national and international economic journals. He has taught in universities all over the world and, during his career, has received a number of scientific distinctions. On account of his broad experience, he is often asked to advise Governments and international organisations.

Sir James A. Mirrlees

Nobel Laureate in Economic Sciences, Professor Sir James A. Mirrlees was appointed as Master of Morningside College, The Chinese University of Hong Kong in August 2009. A pioneer in optimal taxation theory, Professor Sir James Mirrlees was awarded the Nobel Memorial Prize in Economic Sciences in 1996 in recognition of his fundamental contributions to the economic theory of incentives under asymmetric information. He was knighted in 1997. After graduating from the University of Edinburgh in 1957, Professor Mirrlees was admitted to Trinity College at Cambridge University and received his Ph.D. in Economics in 1963. From 1968 to 1995 he was Edgeworth Professor of Economics at the University of Oxford and a fellow of Nuffield College. From 1995 to 2003, he served as professor of political economy at the University of Cambridge. He has been distinguished professor-at-large at CUHK since 2002. Professor Mirrlees has also held visiting professorships at MIT, UC Berkeley, Yale and Melbourne.

Jean-David Naudet[82]

Jean-David Naudet is both statistician and economist. He worked for ten years in the Research Department of AFD (*Agence Française de Développement*), where, amongst other activities, he headed the Evaluation Unit. He recently joined the Africa Department of AFD. Beforehand, he held several positions in the development aid sector in various international organisations, where he worked as programme manager, economist, consultant and researcher. He is the author of numerous publications on evaluation and aid effectiveness.

[82] Jean-David Naudet's article was co-authored by Tanguy Bernard and Jocelyne Delarue, economists, AFD.

Jodi Nelson

Jodi Nelson, director of Strategy, Measurement & Evaluation at the Bill & Melinda Gates Foundation, leads a central team that creates the guidelines and standards for strategy and measurement at the foundation. SME works across the foundation's three programme areas (Global Health, Global Development, and U.S. Programs) to manage a strategy process that helps programme teams and leadership execute on the foundation's goals. SME also spearheads the foundation's commitment to improve measurement and evaluation as a means for learning, decision-making and feedback on the foundation's performance. As a cross foundation team, SME supports best practice in strategy and measurement by facilitating learning across teams and from outside the foundation.

Prior to joining the foundation, Jodi spent eight years at the International Rescue Committee (IRC), where she led an initiative to bridge the gap between academics and aid workers to strengthen the quality of data used to effect and measure change in post-conflict countries. She founded the IRC's department of Research, Evaluation and Learning. Prior to that, Jodi worked with the Asia Society, the Committee for Economic Development, the World Resources Institute, and the Society for International Development.

Jodi has a Ph.D. in political science from Columbia University and has taught at Princeton and New York University.

Catherine Paradeise

Catherine Paradeise was trained in social sciences in an interdisciplinary environment in France and the U.S. She has taught sociology in several universities and *Grandes Ecoles* in France and Canada. She was deputy director of the Department of Social Sciences and Humanities at the French National Centre of Scientific Research (CNRS) (1991-1994), and the head of the Department of Social Sciences (1994-2000) before becoming deputy director of the *Ecole normale supérieure* (ENS) in Cachan (2000-2003). She has written and taught extensively on labour markets, industrial relations, professions and occupations. More recently, she turned to the analysis of organisational issues and public policies in research and higher education institutions. She is presently full professor at University Paris Est, senior researcher at LATTS (*Laboratoire territoires, techniques et sociétés*), a French joint research centre of the CNRS and the University Paris Est. She chairs IFRIS (*Institut Francilien Recherche Innovation Société*), a network institute dedicated to the study of sciences in society, which has recently been awarded several large grants and is certified as a French "Laboratory of Excellence". She is currently senior editor for *Organisation Studies* and *Sociologie du Travail*.

Ruerd Ruben

Ruerd Ruben (1954) holds a Ph.D. in development economics from the Free University Amsterdam. He lived and worked for 14 years in several Central American countries (Nicaragua, Honduras, Costa Rica, El Salvador) and was engaged in programmes for land reform, cooperative development and smallholder agriculture. Thereafter, he was appointed by Wageningen University to coordinate a multidisciplinary research and training programme

on food security and sustainable land use in sub-Saharan countries (Mali, Burkina Faso, Kenya, Ethiopia). He started an innovative programme on the prospects for integrating smallholders into tropical food value chains. In 2006, he obtained the chair in development studies at Radboud University Nijmegen to conduct further research on voluntary organisations and the impact of fair-trade value chains. Since 2010, he has also served as the director of the independent Policy and Operations Evaluation Department (IOB) at the Netherlands Ministry of Foreign Affairs.

Leonce Ndikumana

Léonce Ndikumana is a professor of economics and director of the African Policy Program at the Political Economy Research Institute at the University of Massachusetts, Amherst. He served as director of operational policies and director of research at the African Development Bank, chief of macroeconomic analysis at the United Nations Economic Commission for Africa (UNECA), and visiting professor at the University of Cape Town. He is an honorary professor of economics at the University of Stellenbosch. He has contributed to various areas of research and policy analysis on African countries, including the issues of external debt and capital flight, financial markets and growth, macroeconomic policies for growth and employment, and the economics of conflict and civil wars in Africa. He is co-author of *Africa's Odious Debt: How Foreign Loans and Capital Flight Bled a Continent*, in addition to dozens of academic articles and book chapters on African development and macroeconomics. He is a graduate of the University of Burundi and received his doctorate from Washington University in St. Louis, Missouri.

Miguel Szekely

Dr Miguel Székely is director of the Institute for Innovation in Education, at the *Tecnológico de Monterrey*, in Mexico. Between 2006 and January 2010 he was Under Secretary for Middle Education under the Felipe Calderon Administration. Between 2002 and 2006, he served as Under Secretary for Planning and Evaluation at the Ministry of Social Development, under the V. Fox Administration. He worked as chief of the Office of Regional Development at the Office of the President of Mexico during 2001, as research economist at the Inter American Development Bank from 1996 to 2001, and as researcher in the Economics Department at *El Colegio de México* between 1989 and 1993. He has a Ph.D. in economics and a Masters in economics for development from the University of Oxford, as well as a Masters in public policy and BA in economics from ITAM, Mexico. He has lectured on development economics for Latin America at *El Colegio de México*, ITAM, and the University of Oxford. His is a specialist in education and social policy for Mexico and Latin America, and has researched widely on the topics of inequality, poverty and education. He has 71 academic publications including 8 books, 24 refereed articles in academic journals, and 38 chapters in edited volumes.

Michael Clemens

Michael Clemens is a senior fellow and research manager at the Center for Global Development, where he studies impact evaluation and international migration. Clemens joined the Center after completing his Ph.D. in Economics at Harvard, where his fields were economic development and public finance, and he wrote his dissertation in economic history. His earlier writings have focused on the effects of foreign aid, determinants of capital flows and the effects of tariff policy in the 19th century and the historical determinants of school system expansion. Clemens has been a visiting scholar at New York University, an affiliated associate professor of Public Policy at Georgetown University, and a consultant for the World Bank, Bain & Co., the Environmental Defense Fund, and the United Nations Development Programme. He has lived and worked in Colombia, Brazil and Turkey.

What is AFD ?

Agence Française de Développement (AFD) is a public development finance institution that has been working to fight poverty and foster economic growth in developing countries and the French Overseas Communities for seventy years. It executes the policy defined by the French Government.

AFD is present on four continents where it has an international network of seventy agencies and representation offices, including nine in the French Overseas Communities and one in Brussels. It finances and supports projects that improve people's living conditions, promote economic growth and protect the planet, such as schooling for children, maternal health, support for farmers and small businesses, water supply, tropical forest preservation, and the fight against climate change.

In 2011, AFD approved nearly €6.9 billion to finance activities in developing countries and the French Overseas Communities. The funds will help get 4 million children into primary school and 2 million into secondary school; they will also improve drinking water supply for 1.53 million people. Energy efficiency projects financed by AFD in 2011 will save nearly 3.8 million tons of carbon dioxide emissions annually.

www.afd.fr

Agence Française de Développement
5, rue Roland Barthes – 75598 Paris cedex 12
Tel.: 33 (1) 53 44 31 31 – **www.afd.fr**

ISSN: 2118-3872

Evaluation and its Discontents: Do We Learn from Experience in Development?

The *Agence Française de Développement* and EUDN (European Development Research Network) have been co-organising an annual conference on development since 2003. Over the years, this conference has become a landmark event in Europe for the development community. The March 26 2012 conference gathered over 1,000 participants from more than thirty countries. It challenged a central issue: can we learn from experience in the field of development? If so, how can evaluation contribute and how is it that we seem unable to translate these experiences into practice?

Sir James A. Mirrlees (University of Cambridge, Chinese University of Hong Kong), Paul Gertler (University of California, Berkeley), Jean David Naudet (*Agence Française de Développement*), Jodi Nelson (Bill & Melinda Gates Foundation), Catherine Paradeise (University of Marne-la-Vallée), Ruerd Ruben (Ministry of Foreign Affairs, the Netherlands), Leonce Ndikumana (University of Massachusetts Amherst), Miguel Szekely (former Under Secretary for Planning and Evaluation at the Ministry of Social Development, Mexico) and Michael Clemens (Center for Global Development, Washington) commented the extent to which evaluation methods and approaches can usefully support the learning process in development. They presented facts, questions and reflections to identify how evaluation could, in the long run, contribute to experience-based and better development strategies. Francois Bourguignon (Paris School of Economics), Pierre Jacquet (*Agence Française de Développement*), Mamadou Diouf (Columbia University) and Jean Philippe Platteau (University of Namur) also participated actively in the debates.

Conférences & Séminaires

EUDN

afd AGENCE FRANÇAISE DE DÉVELOPPEMENT