

Challenges and Opportunities of Evaluation in Fostering Development in Sub-Saharan Africa

Albert-Enéas Gakusi

Principal Evaluation Officer, Operations Evaluation Department (OPEV), African
Development Bank. Email: a.gakusi@afdb.org

Alice N. Sindzingre

Research Fellow, National Centre for Scientific Research (CNRS, Paris)-EconomiX,
University Paris X; Visiting Lecturer, School of Oriental and African Studies (SOAS,
University of London, department of economics). Email: sindzingre@wanadoo.fr

**African Economic Conference “Opportunities and Challenges of Development for
Africa in the Global Arena”**

African Development Bank and Economic Commission for Africa

Addis Ababa, 15-17 November 2007

Challenges and Opportunities of Evaluation in Fostering Development in Sub-Saharan Africa

Albert-Enéas Gakusi

Principal Evaluation Officer, Operations Evaluation Department (OPEV), African Development Bank. Email: a.gakusi@afdb.org

Alice N. Sindzingre

Research Fellow, National Centre for Scientific Research (CNRS, Paris)-EconomiX, University Paris X; Visiting Lecturer, School of Oriental and African Studies (SOAS, University of London, department of economics). Email: sindzingre@wanadoo.fr

**African Economic Conference “Opportunities and Challenges of Development for Africa in the Global Arena”
African Development Bank and Economic Commission for Africa
Addis Ababa, 15-17 November 2007**

Abstract

In developing countries, evaluation is a recent but growing practice. It has emerged as a key tool for assessing and measuring outcomes and impacts in order to inform policy makers and donors about the effectiveness of projects, programs and policies. Evaluation is particularly crucial in Sub-Saharan Africa, because of the disappointing economic performances over the past decades of most countries and the mixed effects of aid. The paper critically assesses the strengths and weaknesses of evaluation with respect to its main objectives, i.e. improving the effectiveness of development projects, programs and policies, producing and improving knowledge, and enhancing accountability of policy-makers and donor agencies. The paper contends that evaluations may exhibit limited effectiveness in many situations, because of different factors, such as the weak relevance of evaluation findings, weaknesses in implementing lessons learnt and recommendations when they are relevant, problematic methods used to gather evidence, and issues regarding the access to relevant information. It argues that the objectives of evaluation are constrained by political economy processes, some of them being embedded in aid practices. These processes generate paradoxes regarding the credibility and effectiveness of evaluation: evaluations that are driven by donors and policy-makers may be well-informed and relevant to them, but may be less credible, as they may reflect their interests, which may not be congruent with those of the beneficiaries. Symmetrically, evaluations that are conducted by independent agencies may be more credible in terms of accuracy and conceptual depth: however, they may exhibit information failures, be irrelevant to donors' and policy-makers' interests and be powerless and ineffective in implementing changes in policies. These constraints on evaluation are examined via an analytical framework relying on the economic theories that have long analyzed the concepts of credibility and independence. Despite these limitations, evaluation is still a useful tool, especially when it appropriately and rigorously documents facts. Evaluation results may be useful, but not necessarily for their intended objectives: in particular, they may contribute to the development of democratic institutions. Evaluation effectiveness can be improved, in particular in relying and enhancing local capacities.

Introduction¹

In developing countries, evaluation in its modern definition is a recent but growing practice. The roots of evaluation discipline are usually traced to the 1960s, the era of the War on Poverty and the Great Society in the United States, with its programs aiming at improving education, health and housing, among many other areas. They are also traced to what the evaluation discipline has referred to as Donald Campbell's 'Experimenting Society' (Campbell, 1971), social experiments and experimental studies aiming at assessing the effectiveness of policy interventions (Caracelli, 2000; Campbell, 1976). Evaluation has emerged as a key tool for assessing and measuring outcomes and impacts in order to inform policy makers and donors about the effectiveness of projects, programs and policies.

The role of evaluation is particularly crucial in Sub-Saharan Africa, because of the disappointing economic performances over the past decades of most countries and the mixed effects of aid in its different forms, such as project/program credits or policy-based loans and grants.

The paper critically assesses the strengths and weaknesses of the practice of evaluation with respect to its main objectives, i.e. improving the effectiveness of development projects, programs and policies, producing and improving knowledge, and enhancing the accountability of policy makers and donor agencies. Its arguments are primarily analytical and are substantiated by empirical examples from policy formulation, project and program implementation in Sub-Saharan Africa.

The paper contends that evaluations may exhibit limited effectiveness in many situations because of a series of factors, which include weak relevance of evaluation findings, weaknesses in implementing lessons learnt and recommendations when they are relevant, problematic methods used to gather evidence, and issues in accessing relevant information.

One of its key points is that the objectives of evaluation are constrained by political economy processes, some of them being embedded in aid practices. Evaluation is indeed subject to the limitations that have been recently underscored by a vast literature regarding aid effectiveness, in particular in Sub-Saharan Africa. These political economy processes generate paradoxes and dilemmas regarding the credibility and effectiveness of evaluation: evaluations that are driven by donors and policy-makers may be well-informed and more relevant to them, but may be less credible, as they may reflect their interests, which may not be congruent with those of the beneficiaries. Symmetrically, evaluations that are conducted by independent agencies may be more credible in terms of accuracy and conceptual depth: however, they may exhibit information failures, be irrelevant to donors' and policy-makers' interests and be powerless and ineffective in implementing changes in policies and practices.

These constraints on evaluation are examined via an analytical framework relying on the economic and political economy theories that have long analyzed the concepts of credibility and independence, e.g., the question as to 'who can supervise the supervisor',

¹ The authors are grateful to Raymond Toye for his very useful comments and revisions of this paper, though the usual caveat applies.

which have explored concepts such as incentives for collusion, commitment problems and policy credibility. These theories have highlighted the paradox inherent in credible policies and independent agencies: governments and policymakers cannot credibly commit because there is no supra- or outside entity that has the coercive capacity to bind their actions and enforce their promises - enforcement, however, may be enhanced by other mechanisms, such as reputation: a politician's attention to his reputation may reduce his incentive to renege on his promises. This inherent difficulty to credibly commit has justified the existence of independent agencies. Independent agencies (if no decision-making is delegated), however, may lack the political power and capacity for policymaking that makes their assessments effective. The effectiveness of evaluations may be analyzed according to these concepts, as it is constrained by similar paradoxes and dilemmas.

Despite these limitations, the paper argues that evaluation is a useful tool, especially when it appropriately and rigorously documents facts. Evaluation results are useful, but not necessarily because of their intended objectives. Moreover, their relevance may be acknowledged with a certain time lag. In particular, evaluations may have important effects that go beyond their intended objectives because they may contribute to the emergence and development of democratic institutions.

The paper is structured as follows. Section 1 presents the key debates regarding the evolution of evaluation as a discipline. Section 2 examines a series of determinants of its limited effectiveness. Section 3 elaborates an analytical framework that explores the intrinsic constraints and paradoxes of evaluation, its different degrees of effectiveness resulting from differences in terms of credibility, independence, information and reputation. Section 4 uses this framework in order to explain why evaluation has induced changes in policies in some cases but not in others. Section 5 concludes in underscoring that evaluation exhibits many other effects beyond its intended objectives, and that current evaluation methods can be improved, in particular in enhancing their reliance on local knowledge and skills.

1. The emergence and evolution of evaluation as a discipline

Though it is recent, evaluation of development programs and projects is a growing and increasingly visible practice in developing countries. Under the expression of 'monitoring and evaluation', it now constitutes an autonomous discipline, with professional associations, conferences and publications. Most government agencies and multilateral and bilateral donors now include departments devoted to evaluations, while a large number of private consultancies, firms, think tanks, non-governmental organizations and academic research centers also implement evaluations. Evaluation is now an essential instrument for assessing and measuring outcomes and impacts in order to inform policymakers and donors about the effectiveness of projects, programs and policies.

The discipline has emerged in its modern form in the 1960s, in particular thanks to the works of Donald Campbell on the 'Experimenting Society' (Campbell, 1971; 1976). Since the 1970s, historical events have influenced the evaluation profession and expanded its notions of the use of an evaluation (Caracelli, 2000; Feinstein, 2002). The

notion of use is indeed one of the defining goals of evaluation, which remains a matter of debate, as it may reduce evaluation to mere instruments for social improvements (Henry, 2000). Sophisticated theories of evaluation have been progressively elaborated, which analyzed the conditions of program implementation and mechanisms that mediate between processes and outcomes, with the objective of understanding the rationales underpinning why a program works or does not work (Weiss, 1997; Rogers, 2007).

Conceptual frameworks that evolved over time

The discipline has devised many concepts over time, while conceptual frameworks and evaluation criteria evolved with development paradigms. A series of new concepts came progressively to the fore, such as, among many others, adaptation and adaptive management, capacities, learning, process monitoring, responsiveness, accountability, participatory processes, which involved a plurality of stakeholders such as donors, local implementers, and the target community. Participation, governance, accountability, ownership, actor-oriented approach, are typically concepts built by the development paradigm that has emerged in the 1990s – or “accountability for learning”, “transparency as a mean of public accountability” (Engel et al., 2007). These have become important concepts for the design, implementation and evaluation of development programs.

These theories of evaluation take into account broader dimensions of outcomes, such as more comprehensive determinants of human behavior – e.g., the human motives underpinning individual behavior, the understanding of commitments, of learning processes, and what is coined a ‘constructivist’ approach of development work. In the same vein, theories of evaluation have extended their functions: its role has been increasingly viewed as an instrument that may facilitate “transformative learning in organizations”, or constructivist learning (Preskill and Torres, 2000).

This evolution of objectives towards learning and accountability has been made explicit, for example, in the mandate of the Independent Evaluation Group (IEG) of the World Bank: “the goals of evaluation are to learn from experience, to provide an objective basis for assessing the results of the Bank's work, and to provide accountability in the achievement of its objectives. It also improves Bank work by identifying and disseminating the lessons learnt from experience and by framing recommendations drawn from evaluation findings”.

These perspectives have emerged as especially relevant in developing countries, and theories of evaluation followed the evolution of the perception of the key characteristics of developing countries, in particular its focus on accountability and empowerment. Since the 1990s, low-income countries are indeed often portrayed as characterized by weak governments and institutions, low quality of governance, weak political commitment, lack of accountability and transparency and low level of human capital – education, knowledge and skills.

Likewise, as is well-known, poverty reduction has become a central paradigm in development since the 2000s, as well as the concept of the multidimensionality of poverty, including not only income poverty but also lack of education, health, participation in a society's activities, empowerment and democracy. The concept of multidimensionality has been promoted in academic research by, among others, the

work of Amartya Sen, while its dissemination in donor agencies has been fostered by Sen's advising activities, e.g. at the World Bank or the UNDP. Empowerment has thus become an important notion in donors' practice as well as development research: evaluations of programs now assess not only their impact on poverty with traditional quantitative tools – e.g., cost-effectiveness, benefit incidence, impact analysis, micro-simulations, etc. (Bourguignon and Pereira da Silva, 2003), but also in terms of empowerment (Essama-Nssah, 2002).

The impact of the process of evaluation goes beyond the narrow effects of the project and its immediate neighborhood. The very fact that an evaluation takes place enhances the accountability of decision-makers vis-à-vis the society, as well as democratization processes: as mentioned by Conner (2007), “internationally, evaluation is at the heart of modern developments in governance and democracy”.

Monitoring and evaluation

The evolution of conceptual frameworks and criteria for assessing and evaluating outcomes and impacts has been accompanied by an evolution of reflections and techniques regarding their measurement. Monitoring and evaluation have been progressively distinguished: monitoring refers to assessments of progress *vis-à-vis* pre-established targets, while evaluation refers to broader questions, such as the determinants of success or failure of a policy, program or project, their relevance and the changes that may improve them.

A considerable body of literature has developed that aims at enhancing the monitoring and evaluation (M&E) systems and capacities, in order to increase their performance, such as improving policy making, budget decision making, management, and accountability: among key improvements, they must be ‘evidence-based’ or ‘results-based’. Likewise, monitoring and evaluation refers not only to external entities, as in the case of donors that evaluate a program they finance in a developing country, e.g. via official development assistance, but to governments and their various agencies: donors, governments, private foundations, all may include departments that evaluate their policy- and budget decisions-making, which exhibit various degree of autonomy.

Recent advances regarding M&E systems are summarized in Mackay (2007), who underscores that M&E systems and capacities strongly depend on the level of development and the quality of the functioning of the government and institutions in the countries involved: M&E systems are efficient in developed and middle-income countries, where governments are well-functioning with democratic and accountable institutions.

This may be not the case in low-income countries, such as in Sub-Saharan Africa (SSA), where governments are often weak or non-democratic, policies lack credibility, social tensions and high inequality prevail, civil services and service delivery are hampered by capacity constraints. In these environments, government M&E systems may be weak, as may also be the case regarding the government's demand for M&E. In low-income countries, tight public budgets limit government spending on evaluations.

The limited capacities that characterize many low-income countries underscore their dilemma of choosing between implementing evaluations by external agencies vs. local ones: the latter may not have sufficient capacities, or may lack the necessary freedom

because of authoritarian regimes, interest groups and clientelist functioning. At the same time the current paradigm of ownership may argue against the use of external agencies. Moreover, even if an evaluation is achieved, in the context of weak institutional capacities, the process of M&E may remain ineffective, e.g., when it is not used in the budget and policy process and does not give rise to subsequent reforms.

A crucial issue: measurement

A recurrent question of evaluation theory is that of rigorous measurement. Many evaluations are said to be affected by methodological flaws, for example attributing impacts to a program when they are in fact due to something else – the ‘evaluation gap’ (Center for Global Development, 2006).

Many instruments have been elaborated in addition to more conventional tools for measuring results and classical quantitative indicators: in the same movement as evaluation has come to be more comprehensive and involve multiple actors and factors, as well as considering cumulative and unexpected processes (e.g. non-linearities), the notion of ‘results’ has become more complex, and so, therefore, has the measurement of these results. Various techniques have been elaborated, such as, among others, ‘outcome mapping’, ‘participatory evaluations’, and so on. The question of the number of performance indicators that should be collected is also a matter of debate (Mackay, 2007). Another difficult issue is that of counterfactuals, i.e. evaluation should be able to assess a program and project in taking into account what would have happened in their absence.

These limitations of conventional measurement have given rise to techniques claiming to be more rigorous, such as randomized evaluations. As underscored by Duflo (2004a, b), traditional methods of measuring program impact may be biased due to omitted variables and selection bias, problems that may be solved with the use of randomized evaluations. Duflo and Kremer (2003) define randomized evaluations as evaluations that use control groups: comparison groups are selected randomly from a potential population of participants (e.g., individuals, communities, classrooms and so on) and show their effectiveness on examples of educational programs.

Methods based on randomization produce results that are more precise than those obtained by previous evaluations, and they may significantly re-orient previous policies and projects. On the case of the impact of deworming programs in Kenyan schools, Miguel and Kremer (2003) thus show that there are large differences between the estimates of the social effect that influence attitudes vis-à-vis deworming drugs when they rely on experimental methods (negative estimates) or non-experimental ones (positive estimates). Likewise, a randomized evaluation of the impact of the providing of textbooks to rural primary schools in Kenya shows that the results of previous evaluations – a large positive impact – should be disaggregated, and that it was in fact beneficial only to pupils who already had some academic achievement (Glewwe et al., 2007).

2. Understanding the limited effectiveness of evaluation in Sub-Saharan Africa

An evaluation has obviously, among other objectives, the objective of being used. There can be a great variety of users, such as the policy-makers and agencies that commission it, the direct beneficiaries of a given project, program or policy, the broader groups that are indirectly affected, academic research, and so on. Potential users can include very large groups, such as civil society in the country concerned, policy-makers and civil societies in other countries, international organizations and the like. Users extend far beyond national borders with the globalization of information, the emergence of the global public goods paradigm, and reflections on development that are increasingly conceived at a global scale – e.g., on infrastructure, governance, poverty, service delivery, and so on.

The process of evaluation thus covers a wide range of potential uses and users, from the most proximate and concrete to users at a global scale when evaluations are used for enhancing reflections on the best paradigms of development policies. In assessing successes and failures, evaluations both reflect a particular conceptual framework regarding development, as this framework provides the criteria of the evaluation - e.g. performances in terms of better indicators, accountability, governance, participation, cost-effectiveness, appropriate targeting, and so on -, and contribute to the transformation of this conceptual framework, as they reveal what works and does not work, thus paving the way for different projects and policies.

Evaluations may be effective

Many evaluations have improved the performance and impacts of development policies and programs. A well-known evaluation, for example, has been the assessment by economists from UNICEF of stabilization and adjustment policies in the 1980s (Cornia et al., 1987), which underscored the existence of vulnerable groups that should be protected and their negative impacts on living standards at the household level, health and education. It had an important impact and led the international financial institutions to take into account the social impact of the reforms that conditioned their lending. It contributed to the launch of new projects and surveys (such as the World Bank's 'social dimension of adjustment') and *in fine* to a change of the development paradigm over time and its reorientation on social issues at the end of the 1990s.

As highlighted by the World Bank Operations Evaluations Department (World Bank-OED, 2004), certain evaluations have been influential and have significantly modified previous policies and conceptions of successful development projects. An example is the "citizen report cards" evaluation, which has been promoted in the early 1990s by a local NGO in Bangalore in order to document the views of users regarding the provision and providers of public services (health, transport, water, and the like). It revealed that users had very low level of satisfaction - only 10% of households were satisfied. Follow-up reports have emphasized that this evaluation subsequently encouraged public awareness of the poor quality of public services and incited government departments to improve public service delivery.

Another evaluation that had a positive impact has been the tracking of expenditure and leakages of public funds in primary education services in Uganda. Interestingly, the process relied on the constant exchange between policymaking, ideas and facts in order

to improve a given program: it has been launched in the mid-1990s by a donor, the World Bank, and then continued by academic research (Reinikka and Svensson, 2004), which in turn reinforced the donor program via its results and precision, which finally contributed to disseminate a new broader conceptual framework: shifting the focus not only onto revenues and spending, and onto the amount of allocated funds but also to the possible leakages, and devising better tools for assessing public expenditures, their effective impact and the reaching of intended targets in developing countries. The evaluation found that only 13% of earmarked funds actually reached schools over the period 1991-95 - with the remainder evaporating or used for other purposes - and that 20% of funds allocated for teachers' salaries was in fact going to 'ghost workers'.

This evaluation induced concrete changes in government policy: the government made available to the citizens knowledge about the amounts of public spending. The impact has been dramatic, and in 1999, the funds that effectively reached schools improved to 80-90% of funds (World Bank-OED, 2004). It promoted the new instrument of 'public expenditure tracking surveys', which significantly improved the management of public spending and the monitoring of service delivery, which appeared to be useful much beyond Uganda and has been implemented in many developing countries.

These two types of evaluations include as common features the improvement of accountability and transparency in service delivery via a better dissemination and the use of information both by governments and civil societies, which in turn generate feedback effects in enhancing the standards of service delivery (Sundet, 2004).

A key factor limiting evaluation effectiveness: the political context

However, evaluations exhibit limited effectiveness in many situations, and several factors are examined here.

Limits related to features of evaluations themselves

An important factor is the weak relevance of evaluation findings, in particular from the users' perspective, and weak ownership of findings, when the latter do not involve users. Other factors relate to the fact that some evaluations lack concrete and genuine findings that propose clear policy recommendations. Policy-makers may retain only certain elements from an evaluation, e.g. the least challenging. Lessons learnt and recommendations, when they are relevant, are often not implemented.

Additional factors relate to the methods used to gather evidence, and the access to relevant information. The requisites of monitoring via quantitative indicators and following rigid guidelines devised in developed countries hampers the consideration of local contexts, and therefore adaptive and bottom-up approaches, which may be more effective and foster more commitment by local stakeholders, than top-down methodologies.

Even if changes occur within policy and conceptual frameworks, there is an inertia that is intrinsic to any discipline: for example, as argued by Holvoet and Renard (2007) in the case of SSA, the shift in the concepts of development associated with the new poverty reduction policies (e.g., the Poverty Reduction Strategy Papers-PRSPs) has not been associated with significant change in the M&E systems themselves. Such a change would indeed imply deeper changes, in particular regarding the general organization of

aid and some of its dimensions, e.g., the asymmetric relationship between donors and recipient administrations that are characterized by limited capacities in low-income countries.

The time horizon of evaluations devised in developed countries may differ from that of stakeholders in SSA, both in terms of the time necessary for achieving an evaluation and the time frame of its recommendations. In SSA, users may consider that much longer time frames are necessary and that gradual changes, at the margin, may be more successful in local contexts: indeed, as emphasized by the historian Fernand Braudel (1996), there are different paces in historical change, with economic change being the most rapid, change in political institutions being slower and change in social norms requiring a very long time.

Political contexts

Political contexts are a key factor limiting the application of evaluation principles. Principles of participation, for example, are strongly limited in contexts where social norms foster fragmentation and hierarchies between and within groups, as well as segmented markets, which are a characteristic of developing countries (Bardhan and Udry, 1999).

Principles of accountability and transparency are also severely constrained in undemocratic or clientelist contexts: here secrecy prevails and employees in public administrations or projects are indebted to the groups in power. These principles are similarly difficult to implement in contexts of authoritarian regimes or which lack the rule of law: findings may work against local political interests and balances of political relationships; the evaluator has to take risks and may opt for self-censorship.

All evaluations are conducted in a political context. It may happen that governments, policy-makers or donors may themselves commission evaluations primarily for political purposes, without intending to use their results. Evaluations may be instruments that help to stay in power or climb a hierarchy – they may be ‘rosy analyses’, as shown by Record (2007) in the example of Malawi.

As for initiatives in general, the fact that some evaluations receive attention and others don’t depends on their political environment, e.g. support from political leaders, public voicing, and so on, as argued by Shiffman and Smith (2007) on the example of global health initiatives. The ability to complete an evaluation successfully depends greatly on the evaluator’s ability to navigate the political terrain composed of interest groups who may differently benefit from the implications of evaluation findings if they are to be effectively implemented (Grembowski, 2001, p. xvi).

Evaluations are useful only when their results are actually used in some way by decision-makers, policy-makers, or other groups. As the evaluator delivers the evaluation results, he or she becomes the centre of attention and may be the target of political attacks when findings and recommended measures oppose the interests of specific groups, particularly those who hold power. In this context, the evaluator may favor recommendations that everyone agrees with and ignore recommendations that are controversial but have the potential for major program improvements (Grembowski, p. 263).

In some cases, the findings may challenge and even threaten vested interests and interest groups, which may be close to the government: evaluators may be in a dangerous

situation in terms of career and even more if he/she is a native; the contract may be put to an end and they may be expelled from the country if he/she is a foreigner, which are obvious incentives for consensual evaluation, even if it goes against evidence and if it omits blatant mistakes in the program or project. For instance, when the French agronomist Rene Dumont disclosed his findings to some African countries, further visits by him in those countries became problematic.

Rigorous impact evaluation is also limited in political contexts characterized by factionalism, which may be created by ethnicity, regionalism or religion: this may be the case for randomized experiments for example. The prerequisites to randomized experiments are based on the idea that the results of the evaluation are public goods, implying a general societal consensus (Duflo, 2004a). Social action and understanding is socially negotiated and context-dependent, and reform is as much a matter of argumentation and practical reasoning as it is of experimentation and deductive representation. In these contexts, randomization may in fact be a ‘lure’, as argued by Picciotto (2007): the behavior of participants may change because of the experiment itself, and experimental methods can only be used for a limited range of development interventions that are relatively simple and have been specifically designed for that purpose.

As highlighted by Sadoulet in her comments of Duflo (2004) and the latter’s emphasis on the advantages of randomized evaluations, the impact analysis of development projects must include local political processes. Even if many projects now involve decentralized levels of governance, local participation and more accountability, the political economy of project implementation must be endogenized: evaluating the impact of a project without endogenizing the political economy processes that condition outcomes would be ‘naïve’ (Sadoulet, 2004). Any project requires an understanding of local political processes, in particular the clientelism that organizes the appropriation and redistribution of the project’s rents. Sadoulet mentions the example of the project of Bolsa Escola/Bolsa Familia in Brazil, where local political entities (municipalities) have the power to select the beneficiaries: indeed, the outcomes of the program appear to be very heterogeneous, depending on local political relationships.

Another factor: the transaction costs generated by evaluation

In addition, evaluations may generate high transaction costs for local administrations. Well-known examples are the number of missions per year received by ministries or projects – e.g., the ministry of education. These transaction costs are part of a general problem created by donors’ proliferation. As revealed by the World Bank (World Bank-IDA, 2007), there has been a continuous increase over time of the number of donors: the average number of donors per country nearly tripled over the last half century, rising from about 12 in the 1960s to about 33 in the 2001-2005 period, and the number of international organizations, funds and programs is now higher than the number of developing countries they were created to assist.

Many donor reports have underscored these transaction costs over decades, but interestingly, they have induced only very slow change. Indeed, the lack of coordination between donors ensues from the broader political economy of foreign aid and the inherent competition between donors. Channels and donors’ requirements and

administrative procedures are multiple and represent a heavy burden on implementation capacities that are already notoriously weak in low-income countries.

The health sector is an example of the proliferation of aid channels highlighted by IDA, with more than 100 organizations involved. These negative effects are compounded by the fragmentation of aid. An illustration of transaction costs generated by aid is Tanzania, with 700 projects managed by 56 parallel implementation units. Half of all technical assistance is not coordinated with the government and the country received 541 donor missions during 2005, with 17% involving more than one donor.

Another important factor of limited effectiveness is that evaluations are often ignored or their conclusions and recommendations for change appear impossible to be taken into account, for economic or political reasons, because they challenge an entire architecture of policies, donors and research paradigms at a global scale, which for this reason are more likely to be modified gradually and via marginal changes. In SSA, the repetition - over two decades for some countries, since the 1980s - of adjustment and stabilization programs may be an example, despite their mixed results and an abundance of studies, explanations and warnings - what the IMF IEO had coined “repeated lending” and “prolonged users” since the early-1980s (IMF-IEO, 2002).

Evaluations as dimensions of aid

Another key political economy factor limiting evaluation effectiveness is that the objectives of evaluation are constrained by political economy processes, which are embedded in aid practices. Evaluation is often – but not always, as it may be conducted by national governments and local organizations - a segment in the process of official development assistance. It is often implemented and funded by the same multilateral or bilateral agencies that provide aid. Evaluations are therefore subject to the same limitations that have been recently underscored by a large literature regarding aid effectiveness, in particular in Sub-Saharan Africa.

Interestingly, the many evaluations of aid ineffectiveness have often been as ineffective as evaluations in general and had the same lack of impact: in sum, evaluations are therefore ineffective both because they are part of the aid process and its political dimensions and because aid evaluations themselves are often not taken into account. Aid limitations such as lack of coherence and coordination among donors, proliferation of aid channels and fragmentation, have been highlighted for decades in academic studies and donor reports. Other aspects have been recently emphasized, in particular the detrimental effects of aid at the economic, political economy and institutional levels.

At the economic level, the key aid problems that are examined since the 2000s are those of absorption and spending, as underscored by the IMF (Gupta et al., 2006), and the possible Dutch disease effects. The political economy considerations that weaken the impact of aid are well-documented: aid is implemented by organizations that have vested interests in maintaining aid flows; it is an expression of a country’s political power and an instrument of political relationships between countries. Aid generosity is mainly driven by political motives as aid is an intrinsic dimension of developed countries’ foreign policy (Alesina and Dollar, 1998).

Likewise, as emphasized by the theories of public choice, aid is implemented by donors, and may also be analyzed as a bureaucracy, where individuals seek to maximize their

interest, e.g. their career, and which aims at maintaining its own existence. Public choice therefore highlights the well-known paradox that donor agencies have in fact weak incentives to a success of aid goals and the end of aid. This is one of the reasons why evaluations are ignored and lessons are not drawn from past programs and projects. Learning may indeed be irrelevant for aid organizations (Berg, 2000).

At the institutional level, aid may erode institutions. Aid dependence has negative effects on state institutions, a negative fiscal impact on public revenues. It also affects the relationship between the state and the citizens. If governments raise a substantial proportion of their revenues from aid, they are more accountable to donors than citizens, under less pressure to maintain political legitimacy (Moss et al., 2006).

The questionable preeminence of quantitative assessments

These processes are compounded by the issue of criteria of measurement and appropriateness of indicators. The last decades have witnessed a general movement in research in the social sciences that assimilates scientific rigor to mathematical sophistication and equates ‘proof’ with the use of econometrics.

This movement has also affected evaluations, in particular because of its ‘accountability’ and monitoring objectives. Both policy-makers and evaluation agencies have placed central emphasis on the possibility of quantifying all the notions and facts that were assessed. Preeminence is given to criteria that can be quantified and summarized in figures, statistics, and indices. An evaluation which can be summed up in a simple assessment and in a few ‘messages’, that use quantified evidence, has a better cognitive resilience and is more likely to be widely disseminated - the ‘magic of numbers’. It is better communicated and it is better understood by policymakers as it provides clearer recipes for reform.

The format of an evaluation is more ‘audible’ when it is possible to extract from it a message and a straightforward causality such as: « if the country C had implemented the reform R by X%, its growth rate would have increased by Y% ». For example, the well-known World Bank evaluation of aid, *Assessing aid* (World Bank, 1998) typically included this kind of ‘messages’, e.g., “if all the aid to Zambia had gone into productive investment, it would be a rich country today” (i.e. having a per capita income above \$20000) (p. 10). The result is based on the use of specific hypotheses in a two-gap model, but as often, the discussion of econometrics, methodology and intermediary steps is forgotten and it is this easy-to-remember-message, which is in fine remembered by for the public.

This fosters consensus for most economists (viewing the use of econometrics and statistics as a condition for an argument to be valid), operations, and communication departments, which have a preference for results that are measurable, able to be summarized in a few numbers and statistics, and therefore easily ‘communicated’ to a broader public. This has also political implications. Issues presented as simple, easily measured and with simple solutions gain more political support; they are easier to monitor and therefore more able to shape political priorities, as shown by Shiffman and Smith (2007) on the example of fertility rates and vaccines.

In line with standard economics, it is assumed that everything can be measured. However, while many quantitative indicators can be constructed that measure and

monitor the performances and effectiveness of a given program and project, there are many dimensions of effectiveness that cannot be expressed by quantitative indicators: e.g., the psychological adhesion and trust of individuals, the quality of governance, the fact that the project does not create new social divisions, its relevance to the social and political environment, as well as to local institutions and social norms.

It is broader political, institutional and social dimensions, and moreover their transformation over time, which makes a project is internalized and survives after donors' funding and technical assistance have ended. Institutions are complex devices that result themselves from the combination of other institutions. They cannot be divided into discrete units and quantities that would be stable in time and space: what can be quantified are the formal, external, visible, outcomes of a project, but few of the unobservable processes, such as individual expectations about a project or its adequacy to local norms, or the genuine causality linking different statistics (Sindzingre, 2005). Phenomena that involve aggregate behavior and social systems cannot be modeled, as is the case for representative agents, which explains findings' lack of robustness, e.g., based on cross-country regressions or statistical models (Kittel, 2006).

Finally, this quest for quantifiable measurement not only has effects on the quality of evaluation, but, as shown by Frey and Osterloh (2006), on the example of research performance evaluation, the fact that "what is not measured is disregarded" also entails unexpected costs. The individuals who are subject to evaluations may anticipate their content and change their behavior accordingly (and develop counterstrategies that have a cost). They may erode motivations and lock-in individuals in fixed judgments that do not take into account the complexity of their economic, political and social environment.

3. The inherent constraints upon evaluation: an analytical framework

The paradox of credibility and policy effectiveness in economic theory

Economic theory applied to political economy has demonstrated that all political systems are characterized by commitment problems. The problem of commitment has been analyzed by Kydland and Prescott (1977), who revealed the importance of credibility in economic policy - if governments cannot commit themselves credibly, policies may be futile – and the intrinsic 'time inconsistency' of policies. Citizens may not believe in government policies and promises and may anticipate that governments will change their mind and adapt to circumstances, so it is rational for them to ignore government warnings.

As argued by Acemoglu (2003), individuals holding political power "cannot make commitments to bind their future actions because there is no outside agency with the coercive capacity to enforce such arrangements". Governments and policymakers cannot 'credibly commit' because there is no supra-power that can bind their actions and enforce their promises. For Acemoglu, inefficient institutions and policies exist and persist because they serve the interests of politicians or social groups that hold political power.

The implementation of independent agencies has therefore been viewed as a mean to solve credibility and commitment problems of policy-makers - independent central banks being a well-known example. In the same vein, the lack of ‘agencies of restraint’ has been considered as a key cause of inefficient policies and perpetual policy reversals in SSA, e.g., a lack of ‘supervisors of supervisors’ such as independent central banks or judiciary systems (Collier, 1991).

Likewise, policy credibility is enhanced by hand-binding devices, which may be international treaties or agreements with international financial institutions and their conditionalities. This had been argued in particular for SSA government policies and has constituted an additional justification of conditional lending, policies being considered as more credible if they were tied to some supra-entity, international agreement or institutions (such as the IMF) (Rodrik, 1995).

The concept of credibility of a policy reform - or of a government, or any agency that recommends or prescribes a given policy - has also been analyzed via principal-agent theories, or concepts such as incentives to collusion. In any organization or hierarchy there may be incentives to collusion and coalitions between levels – the principal and the agent, the supervisor and the supervisee, the evaluator and the evaluated (among many other studies, Laffont, 2001). There is always a higher level where there is nobody to ‘supervise the supervisor’, who could guarantee his-her decisions are not biased by particular interests: this affects the credibility of any policy. These problems have long been crucial issues in economic theory. For Stiglitz (1999), the credibility failure ensuing from the impossibility to solve the issue of the meta-supervisor explains the failure of reforms in some transition countries in the first years, e.g. Russia, where supervisors have been ‘captured’ by vested interests (Hellman et al., 2000).

Independence, indeed, is not itself without problem: firstly, independent agencies may themselves lack credibility, and lack of supervision may expose them to political influence or corrupt behavior; secondly, independent agencies have no power to enforce their recommendations, and their independence may make their decisions irrelevant or difficult to use by governments in daily policy-making. There is a paradox inherent to independence, where an increase in credibility may imply a decrease in relevance and power to conduct the appropriate policies.

Unsurprisingly, governments exhibit resistance to the creation of independent agencies as well as to accountability and supervision arrangements, as shown by Quintyn et al. (2007) (on the example of legal and/or institutional frameworks for supervision and financial sector supervisors in recent years in 32 countries): indeed, political control mechanisms often undermine the agency’s credibility.

Credibility, independence, relevance: paradoxes inherent to evaluation

An analytical framework is presented here which explains the variations in the effectiveness of evaluation. These reflections on the commitment problem, credibility and independence and their paradoxes are indeed useful for the analysis of evaluation.

Evaluation is also affected by the problems of commitment and credibility. Evaluation may serve particular interests (of politicians, policy-makers, or donors), which are not necessarily congruent with those of beneficiaries; evaluation agencies may be viewed as ‘meta-agencies’ that may similarly be ‘captured’ by groups that have vested interests in

the policies or projects evaluated, and the accuracy of evaluations cannot be submitted endlessly to evaluations of evaluations and so on (as in the ‘supervisors of supervisors’ problem). In addition, evaluation by definition is not a hand-binding device, donors and policy-makers may ignore it - though accepting an evaluation may improve the credibility and reputation of the donor or policy-maker.

Evaluation is subjected to similar paradoxes, and the question of its credibility and that of its recommendations is always at stake. The independence of an evaluating agency allows it to claim higher credibility – conceptual depth, honesty vis-à-vis political interests. A high degree of independence, however, may mean that the evaluating agency is disconnected from the ‘real’ world of policy-making, politics and power relationships, still more if independence implies a lesser access to insider information. Genuinely independent agencies may indeed be small (e.g. a small local NGO) and have lesser capacity to access information. Such an evaluation may therefore become irrelevant to donors’ and policy-makers interests, it may be ignored and lessen its prospects for effectiveness: higher independence and credibility may mean here lesser relevance.

Symmetrically, evaluations may be driven by donors and policymakers. The evaluating agency may be close to the donors or policymakers, it may even be part of the donor or policymaking institution, e.g., an ‘in-house’ agency, involved in its policy and fully acceding to information. Despite their ‘in-house’ dimension, some agencies are formally independent, such as the World Bank’s Independent Evaluation Group (IEG), which reports to the Bank’s member states (the Board of Executive Directors), not its management, or the IMF Independent Evaluation Office (IEO), similarly independent of IMF management and “at arm’s length from the IMF’s Executive Board”.

These agencies close to policy-makers may also be external consultancy firms or think tanks, which are often large and able to have easier access to the relevant information. Even though they are independent, large agencies are constrained by market effects that stem from the latter’s oligopolistic-oligopsonistic structure and the limited number of players, which may influence the content of the evaluation: market structures create incentives for large consultancy firms to maintain the potential buyers of their products. Large international NGOs are also caught in this constraint. Even if these agencies are formally independent, their evaluations suffer problems of credibility, their independence is called into question and they are exposed to the suspicion that they reflect particular interests. Yet these evaluations may be more relevant to donors, e.g. to their procedures, objectives and interests: lesser independence and credibility may mean here higher relevance.

The biases stemming from the quest for reputation and funding

Evaluations are therefore inherently subject to dilemmas and paradoxes because they include different objectives with different degrees of compatibility: in particular, credibility and independence, vs. information content and relevance. This is not a simple dualistic opposition, as two additional features of the evaluation process introduce further complexity: the quest for reputation on the one hand, the need for funds and information on the other.

Firstly, the evaluating agency may wish to maintain its reputation (this may also be the case for the donor or policy-maker). Secondly, evaluating agencies depend on funding, which may itself be independent (e.g., an academic research project) or provided by the entity – donor, government, project- that is evaluated (e.g., an in-house evaluation department or a consultancy firm); also, to optimize the relevance of their evaluation, they depend on acquiring the most precise information, and in particular insider information, which is provided by the evaluated donor or policy-maker, or networks close to them.

The key point here is that the quest for reputation, as well as for funding and information, blurs the dualistic opposition between the two sets, i.e. “credibility-independence” vs. “less independence, better access to information, relevance but lesser credibility”. Indeed, an ‘in-house’ evaluation department has limited problems of funding and insider information, but it may care a lot about its external reputation, e.g. regarding the honesty of its work, which is an incentive for expressing its independence. In contrast, a formally independent agency, e.g., a think tank or NGO, also cares about its reputation, but it may be very dependent on funds provided by the same donors or governments, and on access to insider information, which is an incentive for establishing close relations with them.

The quest for reputation in turn has a significant impact on the effectiveness, credibility, audience, of an evaluation, whether the agency is independent – ‘in-house’ or external -, small or large. The reputation of independence and quality strengthens the capacity for having a ‘voice’, in particular the use of media. The reputation of the evaluating agency impacts on the donor or policy-maker: enhancing its reputation may be a powerful incentive for the donor or policy-maker to take the recommendations of an evaluation into account and to make them effective – an example being the inclusion of large NGOs in the WTO negotiations.

Picciotto (2004) thus emphasizes that it is civil society organizations that sensitized public opinion with respect to the development incoherence of OECD policies and mobilized political support for specific policy reforms, e.g., on European subsidies, or on the necessity to launch a debt relief initiative, or an international trade agreement on generic drugs.

An evaluating agency able to voice and disseminate its findings is not a sufficient factor of evaluation effectiveness, and donors may be indifferent to reputational effects. Their choice of taking findings into account and changing a project or a development paradigm depends on political calculations and trade-offs, on domestic and international political economy and geopolitical power relationships, particularly when key countries are concerned (e.g., large emerging countries, such as China or India), or on the presence of powerful private interests, as it often happens in large infrastructure projects (e.g., dams, pipe-lines, roads and so on).

These constraints are inherent to evaluation, and variations in effectiveness of evaluation depend on combinations and prioritization of all these elements in given situations.

Evaluation, research and ‘evidence’: indirect relationships

The use of research, in particular its ‘big issues’, also has an impact on the relevance and effectiveness of evaluation – on whether it constitutes an incentive for policymakers to implement changes -, while research may be influenced by evaluation results.

The relationships between evaluation, research and policies are indirect, and so are their respective relationships with empirical facts – ‘evidence’. They are made of reciprocal use and exchanges. In theory, research feeds evaluations and provides more rigorous concepts and methods for assessing and measuring evidence, while at the same time evaluation results feed academic research, and contribute to the ‘bridging of policy and research’ (examples being, among others, the Science Policy Research Unit at the University of Sussex, the Global Development Network initiatives, the ODI Research and Policy in Development/RAPID program; Crewe and Young, 2002).

Large donor agencies are focused on projects and policies, but most include research activities, which at a given time promote specific conceptual frameworks and paradigms. The promotion of such concepts – e.g., the centrality of poverty, governance, gender, empowerment, and so on – stems both from the evolution of concepts and techniques (e.g., surveys) that is constitutive of academic research (e.g., surveys) and facts. Donors and policymakers cannot ignore them as the legitimacy of their projects is grounded in the ‘truth’ of research, and some facts may become salient and debated in public opinion at certain periods (such as in the 1990s the mitigated success of structural adjustment programs in SSA).

The point is that large donors have their ‘in-house’ research departments that closely accompany their policy objectives and provide them with conceptual and empirical basis as well as legitimacy. These policy objectives may in turn bias the objectivity of their research - as shown by the external evaluation of World Bank research on issues such as trade openness or aid (Banerjee et al., 2006).

At the same time, large donors influence external academic research because of their weight, in terms of financing, and in the policies in developing countries. They have therefore an influence on the fact that particular views are preeminent at a given time, according to complex feedback processes (on the concept of poverty, Sindzingre, 2004a). Donors simultaneously shape here the ideas and the ‘facts’ that are important: in sum, ‘truth’ (academic research) and ‘relevance’ (policies, operations) reinforce each other (Sindzingre, 2004b).

Similar feedback processes also characterize the relationships between research and evaluation. Many evaluation agencies, such as the independent bodies of large donors (e.g., the IEG or the IEO) include both conceptual work and policymaking, e.g., technical assistance. In turn, research may be fundamental but also focused on policy. Policy research now feeds evaluations and provides tools for assessment and measurement that are viewed as more rigorous. As everything can be evaluated and monitored, however, there is a risk of endless evaluation: research can itself be evaluated with the tools and techniques of monitoring and evaluation (Hovland, 2007).

Research can also be ignored by policymakers, in particular when it is subservient to political choices. As argued by Lomax Cook (2001), policy-making has a tense relationship with the evidence it relies on, because it follows either the rational actor model, where research plays a major role in policies, or the political model, where

research is only a minor input. On two examples of policy-making (welfare reform in the U.S. and criminal victimization of the elderly), she shows that in both cases research played its most important role in defining the problem and its least important in generating policy solutions. In both cases, research interacted with ideology, interests, and other factors, with public opinion also having an influence on research and policy actors.

4. Why do evaluations have the capacity to change policies? Variations and tradeoffs regarding independence, credibility and reputation

This conceptual framework may help to understand the situations where evaluations have constituted the basis for new donors' or governments' policies - knowing that the latter have limited room for maneuver in SSA - and where relevant evaluations have been ignored. This is an important question, which has significant implications for SSA economies.

In this perspective, evaluations may be analyzed in distinguishing two cases: one causing little change in future policies, programs and projects, another causing significant change. Evaluations may be made by independent external agencies or by internal departments. Evaluations may be reinforced by new empirical evidence that confirm them or not, as well as by new research, or evaluations may themselves reinforce new research findings.

Among many other possible examples, the dynamics of these various cases are explored in a stylized way by programs implemented by the international financial institutions and evaluations conducted by their evaluation offices. Indeed, illustrating the principles of transparency and accountability, many documents are now made available to the public.

Evaluations generating little change: credibility problems and political contexts

Aid effectiveness

Examples of evaluations that brought about little change are many. The evaluation of aid effectiveness is a well-known case, in particular its negative dimensions, e.g., the proliferation of donors and problems of coherence. Reports have been innumerable and often repeated year after year, both made internally by the donor agencies (e.g., the EU) or by academics but did not generate dramatic changes in the broad device of aid, even if some changes have been made in the conceptual framework (e.g. program or projects, *ex post* or *ex ante*, conditional or non-conditional, loans or grants, and so on) and policy practices (e.g., budget support) (e.g., Mourmouras and Rangazas, 2006). The general device of aid is highly resilient, as political (domestic and international) and economic interests in its perpetuation constitute a powerful hindrance to deep changes.

Liberalization policies

In SSA, evaluations generated little change in many cases. In retrospect, for example, in the 1980-90s, despite problematic effects of deindustrialization and reinforcement of the specialization of countries in primary commodities export, evaluations that were critical

on trade liberalization policies recommended by the World Bank in SSA over these two decades were not considered, in particular when they were made by outside agencies (e.g., United Nations agencies or NGOs). External evaluations were ignored – e.g., viewed as poorly informed, under political influence, lacking academic standards, unable to devise appropriate modeling, and so on. This is an example where in-house research produced within the Bank has reinforced the lack of awareness both of programs and internal evaluations regarding possible detrimental effects of the standard policy package, as internal research consisted mostly of econometric exercises underscoring the link between trade liberalization and growth (e.g., Dollar and Kraay, 2000).

This research has been later criticized by further research, inside the Bank with the World Development Report 2000 on poverty, as well as academic research, which showed that trade openness may in some conditions have negative effects on low-income countries or specific groups of the poor, thus igniting controversy within the World Bank and the IMF. It has been also criticized for its weak scientific basis by an evaluation of the World Bank research department itself conducted by external academics (Banerjee et al., 2006). Following the move, the independent in-house evaluation offices of the IMF and the World Bank also now underscore that the relationship is not straightforward and that liberalization policies do not always bring about the expected benefits.

This shows that critical assessments are increasingly considered when research findings accumulate, and when these research findings are included in their evaluations by agencies that have more credibility for the international financial institutions than outside entities, such as the IMF and World Bank evaluation offices, because these offices are in-house and well-informed.

In addition, changes in evaluations pose a time consistency problem: memory of past stances and policies impinges on the credibility of current ones. An independent body, e.g. the IMF IEO, may be critical of past IMF programs on the basis of research demonstrating that, for example, trade openness is not always good for growth. This may be appreciated by external academic research or NGOs as it is a sign of honesty and departure from self-justification, and it may be a scientific assessment: but the credibility of the IEO and of its change of diagnosis are not necessarily enhanced, because it has for years supported the opposite policies.

This integration of challenging research into evaluation close to policymakers, however, does not necessarily generate policy change. Evaluations from both ‘independent’ in-house bodies and research departments may be critical of given programs: yet operational departments may not take them into account.

In sum, when in-house independent bodies make critical evaluations regarding their own institution’s policies, they may face difficult alternatives: either at the internal level, being ignored by operational departments, or for external audiences, being assimilated to the donor itself because it has for years implemented the criticized policies, and therefore being viewed by the public as not very credible.

Privatization reforms

These evaluations made by the main donors’ evaluation departments may generate or accompany change, but precisely because they are made within institutions that

previously supported the criticized policies, their credibility and independence remains questioned.

The evolution of a policy reform such as privatization, which had significant consequences in SSA, is an example that may also be analyzed in this perspective. A number of critical evaluations, coupled with empirical evidence (of mitigated success since its launch in the early-1980s, outside a few sectors such as telecommunication) as well as an evolution in economic research, about the role of the state as well as privatization outcomes in industrialized countries, induced changes in the 2000s regarding the conceptual framework and policies - in particular for the main player in SSA, the World Bank.

This had led to a progressive replacement of the conceptual framework and policy practice relying on the limiting of the role of the state, by more cautious views of privatization, such as the unbundling of public services - the state keeps certain assets and provides the regulatory framework, while the service delivery is operated by private entities according to a variety of contracts (e.g. public-private partnerships) (World Bank, 2004; Bortolotti and Perotti, 2007).

The credibility of the World Bank report (2004) recommending these new policies, however, has been questioned, due to the Bank's rigid privatization policies during two decades despite mixed success in SSA (Bayliss and Fine, 2007a,b), and possible vested interests in the sectors concerned (e.g., water, roads).

Evaluations generating significant changes: a lesser importance of independence, the convergence of evaluation, research and policy-making

The 'Berg report'

In contrast, it has happened that some evaluations laid out the foundations for new paradigms and generated significant policy changes. A well-known example of changes generated by evaluations in SSA is the 'Berg report' written for the World Bank by one of its prominent consultants for SSA, Elliot Berg (World Bank, 1981). This report is an example of the importance, for an evaluation to induce change, of being both inside and outside the policymaking agency (e.g., regular consultant), a position of 'insider independence'.

The report's influence has been reinforced by its use of academic research, e.g., the notion of rent-seeking, which was itself developed by scholars who were very close ('in' and 'out') to international financial institutions and donors (e.g., Krueger, 1974). 'Rent-seeking' emerged as a key cause of economic stagnation in SSA, and the report became a complement both at the policy and theoretical level to structural adjustment and stabilization programs in the early-1980s. The 'Berg report' contributed to the emergence in the 1990s of the paradigm of governance ('poor governance', 'postponed adjustment') promoted in academic studies as well as donor evaluations in SSA.

Changes in the assessments of causalities (economic stagnation explained by rent-seeking and poor governance) met a number of opposite evaluations (in particular from UNECA) but were maintained by several 'in-house' reports, such as the influential World Bank report on SSA (1989) based on background papers that typically associated

analyses made both by academics, some having close ties with the World Bank, and by donor agency officials (Mkandawire, 2004).

The shift towards poverty reduction

Another example is the launch at the end of the 1990s of the Poverty Reduction Strategy Papers (PRSPs) and the Poverty Reduction and Growth Facilities (PRGFs) by the World Bank and the IMF. Despite many commonalities with the previous stabilization and adjustment programs, the great number of external and internal evaluations (by academics, United Nations agencies, bilateral agencies, think-tanks, and so on) on their mixed impact on growth in SSA finally had significant effects on the executive levels of the international financial institutions, while academic research in development economics increasingly centered on poverty during the 1990s (Bourguignon et al., 1991; Kanbur and Squire, 1999). Academic thinking regarding the determinants of growth and the role of the state also evolved during that period, significantly promoted by scholars influential both in academia and within the IMF and the World Bank (such as Joseph Stiglitz).

Independent in-house bodies such as the IMF IEO also made public the disappointing effects of stabilization and adjustment programs (IMF-IEO, 2002, 2003) – its evaluations now underscoring the weaknesses of PRSPs and IMF programs. IEO internal evaluations obviously have privileged access to information regarding the background and impact of IMF programs, but even if they are critical, they may always be suspected of bias. The numerous external evaluations of adjustment programs in SSA, which were perhaps more objective, have been viewed as less relevant in the 1980s by international financial institutions, given the overwhelming importance of stabilization and adjustment in their policies and theoretical framework (e.g., the monetary approach of balance of payments, the Polak model; Polak, 1997) at that time, until empirical evidence of mixed impact in SSA became increasingly visible.

All these factors – research, internal and external evaluations, evidence – generated *in fine* a change in donor country policymakers' views regarding the appropriate policies for developing countries. These factors helped put at the forefront the new conceptual frameworks of 'poverty reduction', 'rehabilitation' of the role of the state and more balanced relationships between donors and governments in program design ('ownership'). They were also reinforced by critical evaluations of financing instruments (loans, project aid) regarding their 'ownership' by recipients, which facilitated a shift from project aid to program aid and budget support.

IMF programs

Another example of internal and external evaluations that had significant influence and generated changes in previous program design is that of the IMF PRGF programs' wage bill ceilings. The latter have always been an element of stabilization in the IMF approach, but may show some discrepancy with the official claim of poverty reduction and improvement of health sectors in the context of the tight budgets of low-income countries, especially in SSA.

Changes resulted from the powerful combination of evaluations both made by the IMF-IEO and a high-profile external think tank (the Center for Global Development, CGD). The point is that the latter is 'more independent' than an independent but in-house body such as the IEO, but at the same time its independence is constrained by a competing

objective of reputation: as the other top think tanks, it maintains its high reputation thanks to high quality and insider information, awareness of the key issues that matter for policymakers, its close links with the World Bank, the IMF and other key agencies and use of former staff from these institutions.

The IMF-IEO evaluation of IMF policy and practice of aid to SSA included a negative evaluation of wage bill ceilings imposed during the two decades of stabilization and adjustment programs (IMF-IEO, 2007): wage bill ceilings were not compatible with the other key aspects of PRSPs, i.e. the objectives of minimal social public spending. The evaluation has prompted the IMF to modify this conditionality, on which significantly the IMF had already a few doubts (*IMF Survey*, March 19, 2007). This shows that policy changes are more likely to be induced when evaluations are congruent with questioning among policymakers, themselves generated by facts and new research. However, the credibility of an 'in-house' agency such as the IEO, though it is better informed, remains questioned because of its link with the evaluatee, the IMF.

On the other hand, the IMF wage ceilings and their adverse effects on PRSPs social spending have been also critically evaluated by the CGD (CGD, 2007; Goldsbrough, 2007, on the case of Mozambique, Rwanda, and Zambia). The evaluation has been headed by an ex-staff from the IMF IEO, with the expert group including several ex-staffs from the World Bank, the IMF or the UN. The CGD evaluation brought about a response by the IMF, arguing that wage bill ceilings have not restricted social spending, and that the CGD evaluation did not take into account key criteria for the IMF such as macroeconomic stability.

Interestingly, the IMF simultaneously presented both the CGD and IEO evaluations as confirmations of its own approach (*IMF Survey*, April 11, 2007). For the IMF, the evaluation of its own independent office, the IEO, was right and came at a time where a consensus emerged within research and the IMF itself that this particular aspect of its policy in SSA (wage bill ceilings) was controversial, had undesirable effects and must therefore be abandoned (Fedelino et al., 2006). The IMF argued that the share of PRGF-supported programs with wage bill ceilings has declined from 40% in 2003-05 to 32% in June 2007 (Verhoeven and Segura, 2007), which is a change both in terms of conceptual framework and policy vis-à-vis the stabilization programs practiced in SSA since the 1980s. This underscores that policymakers may keep from an evaluation – internal or external - the conclusions that contribute to consensus-building.

This stylized story highlights two important points. Firstly, *in fine*, the external evaluation (CGD) does not display significantly more credibility than that of the in-house independent office (IEO). Large external evaluation agencies, think-tanks or NGOs, precisely because they try to maintain their leadership in the 'market' of studies and evaluations about development, are characterized by constraints on their formal independence. Indeed, their funding partially depends on donors; they need to secure access to high-quality and insider information and keep close ties with the entity that is evaluated (and may be perceived as 'satellite' agencies). They may therefore be affected by problems of credibility that are similar to those of internal independent evaluation agencies, e.g., the IMF-IEO; their recommendations may even be less independent than the latter, as their access to information is more costly.

Secondly, this highlights that evaluations are taken into account and generate change when the policymaker already experiences a need for change. This also shows the

importance of the sharing of a ‘common knowledge’ and convergence of ideas between evaluating and evaluated agencies.

Evaluations reinforcing changes generated by research

There are cases where evaluations not only introduce changes in subsequent projects and programs, but where evaluations reflect previous shifts in theoretical paradigms, both within academic and donor-driven research. These evaluations use an analytical framework that is already present in the research-based literature of donors. Evaluations in using new concepts from research reinforce these conceptual changes: such cumulative assessments (from research and from evaluation) makes it is very likely that the evaluation’s findings will be viewed as relevant. This is even more likely if the research and the evaluation are produced by formally independent agencies, which enhances their credibility, but that are also internal, which enhances the relevance of their findings.

An example may be the evaluation made by the Independent Evaluation Group (IEG) of the International Finance Corporation (IFC) of IFC's Development Results, which was based on more than 600 IFC supported investment projects evaluated between 1996 and 2006 (IFC-IEG, 2007). The IEG findings underscore the need for tailored and country-specific strategies with greater IFC-World Bank collaboration, and the necessity of putting more emphasis than in the past on issues such as sustainable rural development, with a focus on agribusiness, rural micro finance, and the environment.

These themes interestingly reflect a paradigm that has already become prominent in academic research and policy-oriented research since the early-2000s. Indeed, the paradigm of country-specific strategies and case by case analysis has become a new consensus in development economics research in the 2000s, as a result of several studies that argue, for example, that development is firstly a process of ‘self-discovery’ (Hausmann and Rodrik, 2003).

Another recent example is the promotion of rural development and the policy recommendation of investing in agriculture by the World Bank flagship report both in terms of research and policy, the 2008 World Development Report, after decades of eclipse. Such a report reinforces changes caused by new facts that have affected economies at a global scale, e.g. agribusiness or the environment, which have in turn generated a large amount of new research, which is in turn reflected in new assessments and policy recommendations.

5. Concluding remarks: beyond intended objectives, the many effects of evaluations

This paper has examined the constraints weighing on evaluation effectiveness and its determinants, in particular those stemming from their environment in terms of political economy.

It has revealed a series of paradoxes that are inherent to evaluation practices and argued that these paradoxes explain the variation in the effectiveness of evaluation and its

capacity for changing policies. These inherent tensions and paradoxes have been explored according to an analytical framework that distinguishes between different key features of evaluation and evaluating agencies: credibility, independence, information, reputation. Differences in relevance and effectiveness of evaluations, i.e. the fact that some evaluations induce policy change or not in given situations, ensue from variations in this set of features, which have been examined via a series of stylized examples.

Despite these limitations, it may be argued that evaluation remains a useful tool, especially when it appropriately and rigorously documents facts. An important point is that evaluation results may be useful, but not necessarily for their intended objectives. Their relevance may moreover be acknowledged with a certain time lag. Indeed, evaluations cannot be assimilated only to 'results' or 'evidence': they constitute processes involving many groups and domains – political, economic and social - in a given country. These processes may be learning processes, which are crucial to the 'ownership' of development; they may also be political processes, i.e. processes modifying existing power equilibria. In particular, evaluations may contribute to the emergence and development of democratic institutions.

Evaluations may have important effects that go beyond intended objectives. This has already been emphasized by Hirschman (1967) in his analysis of a series of World Bank projects, where he argues that side-effects may be in fact 'central' inputs in the achievement of the project, and that project success cannot be defined in a clear-cut way nor reduced to economic indicators, which are moreover often constructed within donor agencies. For Hirschman, projects contribute to development not only due to their explicit objectives but when they improve the 'lacks' that characterize developing countries - lack in capital, skills or institutions that foster development. The success of a development project may imply changes in the institutional and social environment, which is also underscored by Stiglitz (1998) regarding broad institutional transformation as a condition of success of economic reforms in transition countries.

Evaluations in SSA have not generated tangible effects, they have been characterized by little coherence and coordination, and may even be sometimes 'brutal'. External evaluations may be useful as they allow for a fresh look: evaluations, however, are caught in a trilemma of knowledge of evaluation, context or theme, evaluators often being experienced in evaluation but having a limited knowledge of the theme or the context, while others know the theme but have little knowledge of evaluation as a discipline.

This paper has identified the series of limits facing evaluation, which is the first step of identifying improvements. There is room for exploring alternative methods, which would improve evaluation credibility and effectiveness: in particular, the enhancing and greater reliance on local capacities, competence and knowledge would represent significant improvements.

Bibliography

- Alesina, Alberto and David Dollar (1998), *Who Gives Foreign Aid to Whom and Why?*, Cambridge MA, NBER working paper 6612 (pub. *Journal of Economic Growth*, vol. 5, n°1, March, pp. 33-63, 2000).
- Acemoglu, Daron (2003), *Why Not A Political Coase Theorem? Social Conflict, Commitment and Politics*, mimeo, Cambridge MA, Massachusetts Institute of Technology (*Journal of Comparative Economics*, vol. 31, December, pp. 620-652, 2003).
- Banerjee, Abhijit, Angus Deaton, Nora Lustig and Ken Rogoff (2006), *An Evaluation of World Bank Research, 1998-2005*, mimeo, Washington D. C., the World Bank.
- Bardhan, Pranab and Christopher Udry (1999), Institutional Economics and the State in Economic Development, in Pranab Bardhan and Christopher Udry, *Development Microeconomics*, Oxford, Oxford University Press.
- Bayliss, Kate and Ben Fine eds. (2007a), *Privatization and Alternative Public Sector Reform in Sub-Saharan Africa: Delivering on Electricity and Water*, Basingstoke, Palgrave MacMillan.
- Bayliss, Kate and Ben Fine (2007b), *Debating the Provision of Basic Utilities in Sub-Saharan Africa: a Response to Nellis*, Brasilia, UNDP, International Poverty Centre, OnePager, April, n°32.
- Berg, Elliot (2000), Why Aren't Aid Organisations Better Learners?, in Jerker Carlsson and Lennart Wohlgemuth eds., *Learning in Development Cooperation*, Stockholm, EGDI.
- Bortolotti, Bernardo and Enrico Perotti (2007), From Government to Regulatory Governance: Privatization and the Residual Role of the State, *World Bank Research Observer*, vol. 22, n°1, Spring, pp. 53-66.
- Bourguignon, François, Jaime de Melo and Christian Morrisson (1991), Poverty and Income Distribution during Adjustment: Issues and Evidence from the OECD Project, *World Development*, vol. 19, n°11, pp. 1485-1508.
- Bourguignon François and Luiz A. Pereira da Silva eds. (2003), *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, Washington D. C., the World Bank and Oxford University Press.
- Braudel, Fernand (1996), *The Mediterranean and the Mediterranean World in the Age of Philip II*, Berkeley, University of California Press (1st ed. 1949).
- Caracelli, Valerie J. (2000), Evaluation Use at the Threshold of the Twenty-First Century. *New Directions for Evaluation 1978-2000*, n°88, Winter, pp. 99-111.
- Campbell, Donald T. (1971/1988), The Experimenting Society, in Donald T. Campbell, *Methodology and Epistemology for Social Sciences: Selected Papers*, Samuel Overman ed., Chicago, University of Chicago Press.
- Campbell, Donald T. (1976), *Assessing the Impact of Planned Social Change*, Hanover NH, Dartmouth College, Public Affairs Center, occasional paper n°8, reprint.
- Center for Global Development (2006), *When Will We Ever Learn? Improving Lives through Impact Evaluation*, Washington D. C., Center for Global Development, Report of the Evaluation Gap Working Group, May.
- Center for Global Development (2007), *Does The IMF Constrain Health Spending in Poor Countries? Evidence and an Agenda for Action*, Washington D. C., Center for Global Development, Report of the Working Group on IMF Programs and Health Spending.

- Collier, Paul (1991), Africa's External Economic Relations, 1960-1990, *African Affairs*, vol. 90, pp. 339-356.
- Conner, Ross (2007), *Evaluation as a Force for Democracy: Lessons from Abroad*, mimeo, Claremont CA, Claremont Graduate University, Professional Development Workshop Series, Keynote Address, 19th August.
- Cornia, Giovanni Andrea, Richard Jolly and Frances Stewart eds. (1987), *Adjustment with a Human Face: vol. I: Protecting the Vulnerable and Promoting Growth*, Oxford, Oxford University Press and UNICEF.
- Crewe, Emma and John Young (2002), *Bridging Research and Policy: Context, Evidence and Links*, London, Overseas Development Institute (ODI), working paper 173.
- Dollar, David and Aart Kraay (2000), *Growth is Good for the Poor*, Washington D. C., the World Bank, policy research working paper 2587.
- Duflo, Esther (2004a), *Evaluating the Impact of Development Aid Program: the Role of Randomized Evaluations*, mimeo, Paris, AFD Conference, 25 November.
- Duflo, Esther (2004b), *Scaling Up and Evaluation*, mimeo, Annual World Bank Conference on Development Economics.
- Duflo, Esther and Michael Kremer (2003), *Use of Randomization in the Evaluation of Development Effectiveness*, mimeo, Washington, D.C., World Bank Operations Evaluation Department (OED) Conference on Evaluation and Development Effectiveness, 15-16 July (pub. in George Pitman, Osvaldo Feinstein and Gregory Ingram eds., *Evaluating Development Effectiveness*, New Brunswick, NJ: Transaction Publishers, 2005).
- Engel, Paul, Niels Keijzer and Charlotte Ørnemark (2007), *Responding to Change: Learning to Adapt in Development Cooperation*, Maastricht, ECDPM Policy Management Brief n°19, March.
- Essama-Nssah, B. (2002), *Empowerment and Poverty-Focused Evaluation*, mimeo, Washington D. C., the World Bank.
- Fedelino, Annalisa, Gerd Schwartz and Marijn Verhoeven (2006), *Aid Scaling Up: Do Wage Bill Ceilings Stand in the Way?*, Washington D. C., International Monetary Fund working paper WP/06/106.
- Feinstein, Osvaldo N. (2002), Use of Evaluations and the Evaluation of their Use, *Evaluation*, vol. 8, n°4, pp. 433-439.
- Frey, Bruno S. and Margit Osterloh (2006), *Evaluations: Hidden Costs, Questionable Benefits, and Superior Alternatives*, mimeo, Zurich, University of Zurich.
- Goldsbrough, David (2007), *Does the IMF Constrain Health Spending in Poor Countries?*, Washington D. C., Center for Global Development Brief, July.
- Grembowski, David E. (2001), *The Practice of Health Program Evaluation*, Thousands Oaks CA, Sage Publications.
- Gupta, Sanjeev, Robert Powell and Yongzheng Yang (2006), *Macroeconomic Challenges of Scaling Up Aid to Africa: a Checklist for Practitioners*, Washington D. C., International Monetary Fund.
- Hausmann, Ricardo and Dani Rodrik (2003), Economic Development as Self-Discovery, *Journal of Development Economics*, vol. 72, n°2, December, pp. 603-633.
- Hellman, Joel S., Geraint Jones and Daniel Kaufmann (2000), "Seize the State, Seize the Day": *State Capture, Corruption and Influence in Transition*, Washington D. C., the World Bank, policy research working paper 2444.

- Henry, Gary T. (2000), Why Not Use?, *New Directions for Evaluation 1978-2000*, n°88, Winter, pp. 85-98.
- Hirschman, Albert O. (1967), *Development Projects Observed*, Washington D. C., Brookings Institution Press.
- Holvoet, N. and Robrecht Renard (2007), Monitoring and Evaluation under the PRSP: Solid Rock or Quicksand?, *Evaluation and Program Planning*, vol. 30, pp. 66–81.
- Hovland, Ingie (2007), *Making a Difference: M&E of Policy Research*, London, Overseas Development Institute (ODI), working paper 281.
- International Finance Corporation (IFC)-Independent Evaluation Group (2007), *Independent Evaluation of IFC's Development Results 2007: Implications from 10 Years of Experience*, Washington D. C., International Finance Corporation.
- International Monetary Fund-Independent Evaluation Office (2002), *Evaluation of Prolonged Use of IMF Resources*, Washington D. C., International Monetary Fund.
- International Monetary Fund-Independent Evaluation Office (2003), *Fiscal Adjustment in IMF-Supported Programs*, Washington D. C., International Monetary Fund.
- International Monetary Fund-Independent Evaluation Office (2007), *The IMF and Aid to Sub-Saharan Africa*, Washington D. C., International Monetary Fund.
- Kanbur, Ravi and Lyn Squire (1999), *The Evolution of Thinking about Poverty: Exploring the Interactions*, mimeo, Washington D. C., the World Bank.
- Kittel, Bernhard (2006), A Crazy Methodology? On the Limits of Macroquantitative Social Science Research, *International Sociology*, vol. 21, n°5, pp. 647-677.
- Glewwe, Paul, Michael Kremer and Sylvie Moulin (2007), *Many Children Left Behind? Textbooks and Test Scores in Kenya*, Cambridge MA, NBER working paper 13300.
- Krueger, Anne O. (1974), The Political Economy of the Rent-Seeking Society, *American Economic Review*, vol. 64, n°3, pp. 291-303.
- Kydland, Finn and Edward Prescott (1977), Rules Rather Than Discretion: the Inconsistency of Optimal Plans, *Journal of Political Economy*, vol. 85, n°3, June, pp. 473-491.
- Laffont, Jean-Jacques (2001), *Enforcement, Regulation and Development*, mimeo, Nairobi, AERC meeting, May.
- Lomax Cook, Fay (2001), *Evidence-based Policy-making in a Democracy: Exploring the Role of Policy Research in Conjunction with Politics and Public Opinion*, mimeo, San Francisco, American Political Science Association Annual Conference, 1st September.
- Mackay, Keith (2007), *How to Build M&E Systems to Support Better Government*, Washington D. C., the World Bank, Independent Evaluation Group (IEG).
- Miguel, Edward and Michael Kremer (2003), *Networks, Social Learning, and Technology Adoption: the Case of Deworming Drugs in Kenya*, mimeo, Berkeley, University of California, Berkeley and Cambridge MA, Harvard University.
- Mkandawire, Thandika (2004), The Itinerary of an Idea, *D+C Development and Cooperation*, vol. 31, n°10, October.
- Moss, Todd, Gunilla Pettersson and Nicolas van de Walle (2006), *A Review Essay on Aid Dependency and State Building in Sub-Saharan Africa: An Aid-Institutions Paradox?*, Washington D. C., Center for Global Development, working paper 74.

- Mourmouras, Alexandros and Peter Rangazas (2006), *Foreign Aid Policy and Sources of Poverty: A Quantitative Framework*, Washington D. C., International Monetary Fund, working paper 06/14.
- Picciotto, Robert (2004), *Policy Coherence and Development Evaluation: Concepts, Issues and Possible Approaches*, Paris, OECD Workshop: Policy Coherence for Development, 18-19 May.
- Picciotto, Robert (2007), *The New Environment for Development Evaluation*, address, Niamey, Fourth Conference of the African Evaluation Association, 19th January.
- Polak, Jacques (1997), *The IMF Monetary Model at Forty*, Washington D. C., International Monetary Fund, working paper WP/97/49.
- Preskill, Hallie and Rosalie Torres (2000), The Learning Dimension of Evaluation Use, *New Directions for Evaluation 1978-2000*, n°88, Winter, pp. 25-37.
- Quintyn, Marc, Silvia Ramirez and Michael W. Taylor (2007), *The Fear of Freedom: Politicians and the Independence and Accountability of Financial Sector Supervisors*, Washington D. C., International Monetary Fund working paper WP/07/25.
- Record, Richard (2007), From Policy to Practice: Changing Government Attitudes towards the Private Sector in Malawi, *Journal of International Development*, vol. 19, pp. 805–816.
- Reinikka, Ritva and Jakob Svensson (2004), Local Capture: Evidence from a Central Government Transfer Program in Uganda, *Quarterly Journal of Economics*, vol. 119, n°2, pp. 678-704.
- Rodrik, Dani (1995), Why is There Multilateral Lending?, in Michael Bruno and Boris Pleskovic eds., *Annual World Bank Conference on Development Economics*, Washington D. C., the World Bank, pp. 167-193.
- Rogers, Patricia J. (2007), Theory-Based Evaluation: Reflections Ten Years On, *New Directions for Evaluation*, n°114, Summer, pp. 63-81.
- Sadoulet, Elisabeth (2004), Comment of “*Evaluating the Impact of Development Aid Programs: the Role of Randomized Evaluations*” by Esther Duflo, mimeo, Paris, AFD conference, 25th November.
- Shiffman, Jeremy and Stephanie Smith (2007), *Generation of Political Priority for Global Health Initiatives: A Framework and Case Study of Maternal Mortality*, Washington D. C., Center for Global Development, working paper 129.
- Sindzingre, Alice (2004a), The Evolution of the Concept of Poverty in Multilateral Financial Institutions: the Case of the World Bank, in Morten Boas and Desmond McNeill eds., *Global Institutions and Development: Framing the World?*, London, Routledge.
- Sindzingre, Alice (2004b), ‘Truth’, ‘Efficiency’, and Multilateral Institutions: a Political Economy of Development Economics, *New Political Economy*, vol. 9, n°2, June, pp. 233-249.
- Sindzingre, Alice (2005), *Institutions and Development: A Theoretical Contribution*, mimeo, The Hague, Institute of Social Studies (ISS), Economic Research Seminar, 28th April.
- Stiglitz, Joseph E. (1998), *Towards a New Paradigm for Development: Strategies, Policies and Processes*, Geneva, UNCTAD, Prebisch Lecture.
- Stiglitz, Joseph E. (1999), *Quis Custodiet Ipsos Custodes? Corporate Governance Failures in the Transition*, Paris, CAE and the World Bank, Annual World Bank Conference in Development Economics.

- Sundet, Geir (2004), *Public Expenditure and Service Delivery Monitoring in Tanzania: Some International Best practices and a Discussion of Present and Planned Tanzanian Initiatives*, Dar-es-Salaam, HakiElimu working paper 04-7.
- Verhoeven, Marijn and Alonso Segura (2007), *IMF Trims Use of Wage Bill Ceilings*, Washington D. C., International Monetary Fund, IMF Survey Magazine, 5th September.
- Weiss, Carol H. (1997), Theory-Based Evaluation: Past, Present, and Future, *New Directions for Program Evaluation*, n°76, pp. 41-55.
- World Bank (1981), *Accelerated Development in Sub-Saharan Africa*, Washington D. C., the World Bank (the 'Berg Report').
- World Bank (1989), *Sub-Saharan Africa: from Crisis to Sustainable Growth: A Long-Term Perspective Study (LTPS)*, Washington D. C., the World Bank.
- World Bank (1998), *Assessing Aid: What Works, What Doesn't and Why*, Washington D. C., the World Bank.
- World Bank (2004), *Reforming Infrastructure: Privatization, Regulation and Competition*, Washington D. C., the World Bank.
- World Bank-IDA 15 (2007), *Aid Architecture: an Overview of the Main Trends in Official Development Assistance Flows*, Washington D. C., the World Bank, International Development Association Resource Mobilization.
- World Bank-OED (2004), *Influential Evaluations: Evaluations that improved Performance and Impacts of Development Programs*, Washington D. C., the World Bank, OED/Operations Evaluation Department.