



Network of Networks on Impact Evaluation

■ DAC Evaluation Network ■ Evaluation Cooperation Group ■ International Organization for Cooperation in Evaluation ■ UN Evaluation Group

Fostering Impact Evaluations at Agence Française de Développement: A Process of In-house Appropriation and Capacity-Building

Jean David Naudet and Jocelyne Delarue

NONIE WORKING PAPER NO. 2

January 2008

What is NONIE?

Nonie is a network of networks for impact evaluation comprised of the DAC Evaluation Network, The United Nations Evaluation Group (UNEG), the Evaluation Cooperation Group (ECG), and a fourth network drawn from the development evaluation associations (AfrEA, IOCE, IDEAS, ReLAC, and IPEN). Its purpose is to foster a program of impact evaluation activities based on a common understanding of the meaning of impact evaluation and approaches to conducting impact evaluation.

To this end a task team has been constituted and tasked with the following activities:

1. Preparation of impact evaluation guidelines
2. Agreeing collaborative arrangements for undertaking impact evaluation, leading to initiation of the program
3. Developing a platform of resources to support impact evaluation by member organizations

NONIE Working Papers

NONIE working papers present conceptual papers and impact evaluation findings. They may have been published elsewhere, e.g. as government or agency reports, but are included in the NONIE series to increase dissemination. Feedback on papers via the NONIE website is welcome.

Fostering Impact Evaluations at Agence Française de Développement: A Process of In-house Appropriation and Capacity-Building

Jean David Naudet and Jocelyne Delarue

1 AFD strategy towards impact evaluation

1.1 The impact issue relates to knowledge production and results-based management

The mandate of Agence Française de Développement (AFD) is to contribute to the financing of economic, social and/or environmental development projects. AFD provides assistance to the public sector (state administrations, public enterprises, and local governments), the private sector, and local associative networks. It offers a large array of financial instruments to help implement sustainable development projects. AFD's goals are to (1) reduce poverty and inequalities by 2015 (MDGs), (2) promote economic growth and (3) protect global public goods (climate, biodiversity, and global health).

Since the beginning of this decade, AFD has been engaged in the renewal of its strategic orientations, taking place in the larger context of the reform of French cooperation. Amongst strategic shifts experienced by AFD, two are noteworthy.

The first shift is the implementation of results-based management, in line with Paris Declaration commitments. As early as 2002, the first AFD strategic plan (POS I) expressed a strong concern about impact. It recommended developing a results-oriented monitoring system, partly based on impact indicators, and being selective in funding on the basis of impact assessments. AFD management places a high priority on ensuring that AFD's assistance is focused on development results and impacts. At all levels, the attention is now on increased and demonstrated effectiveness of development assistance. Management for Development Results is used systematically throughout the project cycle. Aggregated indicators are monitored for expected and real development results. Their definitions are standardised and harmonised with those of international agencies. Monitoring the contribution to the MDGs measures France's and AFD's commitments in terms of resources and results. The economic analysis of development projects goes beyond their financial sustainability. Economic costs and benefits are assessed for the society as a whole, including environmental goods and services. Analysing how each stakeholder group benefits from a project will inform the choice of transfer mechanisms.

The second shift is to a clear focus on knowledge production as a necessary complement of financial activity. According to the Second Strategic Plan (AFD 2007) the focus will be particularly on major ODA-related topics in order to contribute to French policy stances, to participate in partners' capacity-building, and to fuel international debate.

These new orientations have contributed to developing a collective interrogation on the impact issue. One consequence has been growing awareness of the lack of knowledge of AFD's impacts, mentioned as the 'knowledge shortage' in the 'evaluation gap report' published by the Centre for Global Development (CGD) (CGD 2006).

In fact, whereas certain types of projects or financing products, like microfinance or rural roads, have mobilised large amounts of financing, until recently AFD never conducted rigorous research on what actually works and what is attributable to its programmes. The evaluations implemented in the past produced very little information about medium- or long-term effects and none about the net change in outcomes attributable to the projects. Very understandably, project managers focus in the early phases on project design and implementation, and leave the necessary preparations for a sound evaluation for later. Until now, there has been very little incentive to design an appropriate information system linked with the projects in order to assess their impact *ex post*. Nor has there been incentive to compensate for the cost and development time of such a system.

The new AFD approach of strategy-driven operations and of management for development results means that, more and more, operational services must demonstrate impact to obtain funds. AFD are increasingly planning to capitalise on and measure the impact of their operations, but for the moment there is a lack of human resources and budget to implement this across the board. It seems important for AFD to first test whether rigour in impact evaluations, rather than focusing on accountability or process, improves the quality of feedback on operations.

At the same time, the international debate about impact evaluations reveals conceptual, methodological, and practical difficulties. The attribution/contribution question comes under consideration in the implementation of results-based management, particularly for an institution like AFD often involved in co-financing activities. The double meaning of the term impact in the development discourse – either long-term effects or rigorously attributable effects – remains a permanent source of ambiguity. The subtle difference between impact and additionality, a concept more frequently used (also with much ambiguity) for financial mechanisms and particularly as regards climate change, is a concern for an institution divided between the direct financing of public policy in poor countries and participation in financial incentives for global public goods in emerging countries. Methodological questions are directly linked with these conceptual debates: how to measure and attribute long-term effects, how to build a counterfactual, which kind of baseline is necessary, which impact indicators are relevant, etc.

This period of debate is very rich for AFD. It is persuading AFD's staff that a 'one size fits all' conceptual and methodological approach cannot be the unique answer to the shortage of knowledge on impact and that AFD cannot draw much profit from a 'black box' tool for impact measurement. AFD has decided to pursue and diversify methodological and knowledge investment, especially through pilot operations and specialised partnerships, in this field of impact assessment for the forthcoming period and to actively participate in the international debate on this topic.

1.2 Impact evaluations are an important part of the evaluation process

Following these new orientations, AFD has undertaken a reform of its evaluation function since 2006. The reform is based on a two-pillar system: the decentralisation (towards local agencies) of a systematic external evaluation of individual financings and

a reorientation of the Evaluation Unit towards evaluation quality, strategic evaluation, and knowledge production – including impact measurement.

Prior to 2006, the evaluation function had been based mainly on in-house project evaluation, driven and implemented by the Evaluation Unit and placed under the direct authority of the Head of Strategy. About 15 per cent of projects financed by AFD were subjected to final evaluation. Project evaluations were presented and discussed in the Evaluation Committee, chaired by the General Management. The general assessment was that AFD did not make the most of its evaluation system. The Evaluation Unit was isolated from the rest of the institution, feedback was weak and it was difficult to mobilise the Evaluation Committee.

At the beginning of 2006, the Evaluation Unit was integrated as a division of the Research Department (the equivalent of the Knowledge Department in some institutions), which is part of the Strategy Branch. This repositioning demonstrates a clear decision to establish a link between knowledge production and evaluation. Furthermore, it is stipulated in the evaluation principles that part of the evaluation work should be social science research.

The revitalisation of evaluation has been based on four principles:

1. *The sharing of evaluations*: evaluation should not be a restricted activity centralised in a dedicated unit and only devoted to informing top management. Evaluation reports should directly concern and involve all development actors, and in particular operational departments and local partners.
2. *Synergy between evaluation and research*: part of the evaluation work was to be research applied to the analysis of processes and results of projects and programmes.
3. *Focus on capitalisation of experiences and institutional learning*: evaluations directed to feed this process are formative rather than summative, centred not only on policies and strategies, but also on professional practices including ongoing operations.
4. *Gradual mixing of external and internal analyses*: although external evaluation should be the most common way of functioning, internal evaluation remains necessary on the one hand to make the process of institutional learning effective and, on the other hand, to keep know-how up to date inside the Evaluation Unit and make it professionally attractive.

AFD has also recently adopted a number of new evaluation tools: decentralised evaluations, strategic evaluations, thematic capitalisation, and rigorous impact evaluations. Decentralised evaluations will be commissioned by geographical departments and local agencies, shared with local partners, and entrusted to external experts – giving priority to local experts. Strategic evaluations continue to be

commissioned and piloted by the Evaluation Unit, under the initiatives of management and supervisory Ministries. Thematic capitalisations are being developed by the Evaluation Unit based on a comparative analysis of groups of development operations, completed or ongoing, financed by AFD with or without other partners. Rigorous impact evaluations are performed in partnership with specialised academic teams and interested local partners.

Since it was integrated into the Research Department, the Evaluation Unit has received a mandate to develop impact evaluations. The link between knowledge production and evaluation is actually a key factor in facilitating investment in impact evaluation. A significant impediment to the development of impact evaluation inside evaluation units is probably the frequent formal separation between research activities and evaluation activities and the impression that social science research and evaluation are two different jobs.

As mentioned earlier, the AFD impact evaluation programme will be developed progressively, partly implying internal intellectual investment on the impact issue. In broad outline, the programme pursues the following objectives. The first goal is strategic, aiming at producing sound knowledge about what works and what does not in development policies for southern partners, AFD sectoral policymakers, and more generally, for the development community. Second, AFD pursues a methodological goal that will be reached through the development of in-house mastery of different impact measurement tools. The last objective is to build up partnerships through joint knowledge production with different southern partners and active collaboration with specialised scientific teams.

Ultimately, AFD intends to conduct this progressive experience while regularly sharing and enriching it through active participation in international networks on impact evaluations and through the pooling of results and methodologies.

2 A set of experiences

2.1 A preliminary phase

AFD started to foster sound impact evaluations as early as 2003 by financing research on the impact of multi-donor agricultural development projects on farmers' income in Guinea (Delarue, 2007). This impact evaluation was conducted by the Agroparistech and aimed at proposing an adapted methodology for quantifying impact using a qualitative approach. It was carried out entirely *ex post*, and relied on a structured survey using recall to collect data on the pre-intervention period and on in-depth interviews with one hundred farmers. In order to quantify the net change in the farmers' income produced by one of the projects, the researcher identified a set of farmers who were directly or indirectly affected by the project and an unaffected group which comprised a credible comparison group. Two types of projects were evaluated: the development of inland valleys for irrigated rice cultivation and a public agro-industry producing rubber and palm oil.

By conducting an in-depth study of a limited number of production units, the evaluator was able to identify a typology of production systems which existed before the project. In order to set up a counterfactual, a judgement sample was then realised by choosing production units which belonged to the same initial type of production system and which evolved with or without the project.

In-depth understanding of the endogenous and exogenous factors influencing the evolution and possible trajectories of production systems enabled the evaluator to rigorously identify the individuals whose evolution with or without the project were comparable. The evaluator's direct involvement in data collection was essential, hence the importance of a small sample. It would not have been possible to gather reliable data on yields, modifications to production structures over time and producers' strategies from a large survey sample in a rural context.

Then, based on the understanding of the ways projects proceeded and of the trajectories of these farmers, with or without the project, it was possible to build a quantitative model, based on Gittinger's method of economic analysis of development projects (Gittinger 1982). As the initial diversity of production units was well identified before sampling, this model was constructed for each type of farming system existing before the project. Understanding the possible evolutions for each farming system with and without the project allowed for the estimation of the differentials created by the project in farmers' incomes.

Although the objective differences between each production unit studied appear to leave room for the researcher's subjectivity when constructing the typology and sample, the rationale behind the production system concept made it possible to transcend this possible arbitrariness. What underlies this methodological jump from a small number of interviews to a model is the demonstration that a finite number of types of production systems exists in reality.

The primary interest of this new method was to provide the opportunity to build a credible impact assessment entirely *ex post*. Second, it gave an estimate of the impact on different types of farming systems, making explicit the existing inequalities in the distribution of the projects' benefits. Third, it permitted a subtle understanding of the reasons why the desired impacts materialised or did not.

The results from this first impact assessment were available after four years of fieldwork and data treatment. They were presented to the Guinean authorities and to the local representatives of the main donors in the rural sector. In the field, the results were delivered to the local communities interviewed and to the farmers' syndicates. The Minister of Agriculture declared that he would try to foster more impact evaluations on agricultural development projects. Unfortunately, there is little hope that the conclusions of this research will change national policy on these types of projects in the absence of an institutionalised forum for discussing it among the different stakeholders.

2.2 *The second impact assessment financed by AFD*

The next study concerned a microfinance institution called ADéFI, serving micro-entrepreneurs in Antananarivo (Madagascar). It was responding to a request emerging from both the ADéFI management and AFD to produce valid data about the project and to analyse it with a scientifically robust method. The impact evaluation was conducted between 2003 and 2005 by researchers from the Institut de Recherche pur le Développement/Développement Institutions et Analyses de Long terme (IRD-DIAL), a French research centre (Gubert and Roubaud 2005).

At first, the methodology consisted of comparing the situation of a representative sample of micro-enterprise ADéFI clients with a comparison group, constructed through a standard matching technique (propensity score matching). This first quantitative impact evaluation, was a 'post-test project and comparison groups' evaluation design (Bamberger *et al.* 2006). It relied on 255 interviews conducted in 2001 and was at the time complemented by a qualitative analysis based on open interviews with a limited number of ADéFI's clients.

This analysis was quite encouraging concerning the project's target group (in accordance with the project's theory) and in terms of impacts (on the turnover or production of clients' enterprises). But this first evaluation design was not very robust and a second phase was programmed in order to be able to apply a double difference technique. This second phase, consisting of two successive surveys on the same panel of enterprises in 2003 and 2004, enabled the gathering of information on the evolution of the panel subjects and the inclusion of new variables in the matching process which enhanced the quality of results by rendering clients and non-clients even more similar than in phase one.

Unfortunately, the enterprises in the treatment and in the comparison groups of 2001 were not always found again: the attrition rate was respectively 22 per cent and 23 per cent in the two groups in 2003. In 2004, only 55 per cent of the enterprises in the original panel were still active. This low survival rate shows the great fragility of small enterprises, and against expectations, the clients of ADéFI were more affected than non-clients. Whereas 255 enterprises were interviewed in 2001, only 107 interviews could be used for the panel analysis in 2004.

Other methodological aspects were improved during the second phase. The observables selected for the propensity score regression (probit) for both phases included the micro-entrepreneur's gender, age, educational level, type of learning, economic branch of the enterprise, type of premises in which the activity was undertaken, the creation date of the enterprise, the initial workforce, the initial value of capital stock, etc. In 2001, this information was requested concerning the year of the creation of the enterprise, which was not identical for all of them. In 2003 and 2004, this information was asked about 1997 in order to control for the differences in characteristics that prevailed between clients and non-clients at the point in time when ADéFI started.

Several variables of interest relating to the economic results of the micro-enterprises were studied: turnover, production, added value, workforce, capital, and finally, productivity of work and capital. The impact of the microcredit on these variables appears to be positive and statistically significant in 2001 and 2004. But, in general, the impact measured in 2004 appears to be smaller than the one first assessed with the 2001 data. A methodological cause can partially explain this difference: the matching was more rigorous with the 2004 data and there is probably a bias in the 2001 results.

In fact, because the matching in 2001 was based on the characteristics of the enterprises in the year they were created, the propensity score was not based on output variables (as, for instance, turnover, production, etc.). On the contrary, in 2004, the 1997 turnover was included amongst the variables used in the model to predict participation. This contributed to achieving a better match. A simple comparison between the two sets of variables to calculate the propensity score on the basis of 2004 data showed a significant difference in the impacts measured, even if the impact remained positive for all variables. This test demonstrated the extreme importance and difficulty of building an adequate comparison group, particularly when there is no baseline data.

Finally, the use of the double difference technique between 2001 and 2004 gave very different results from the aforementioned single-period measures. With the latter, the project showed a positive impact on productivity and different outputs. In contrast, with double difference, none of the measured impacts was significant, which means that the evolution of the economic results for clients and non-clients is identical on average and that the project did not succeed in activating a growth dynamic for its clients.

This impact evaluation demonstrated how difficult it is to collect panel data on the clients of a microfinance project, because of the high attrition rate in this case. It was linked with the vulnerability of the micro-enterprises and to their propensity to change location, compelling interviewers to track them, often in vain. This evaluation also showed how sensitive the impact measure was to the matching quality.

This study is one of the rare impact evaluations measuring the impacts of an MFI on micro-entrepreneurs. It was also AFD's first completed experience with rigorous impact evaluations, in 2005. The rigour and transparency with which the research team conducted the scientific work helped AFD's institutional learning about conducting impact evaluations. It encouraged AFD to foster new impact studies, addressing the methodological limits of this first exercise, and it led to a larger budget for impact evaluations. In particular, the next impact evaluations were programmed long before the start of the project, in order to gather necessary information about the initial situation through a comprehensive baseline survey.

2.3 A randomised controlled trial of microfinance in Morocco

The first experimental impact evaluation financed by AFD concerns a microfinance institution called Al Amana in rural areas in Morocco.

There still exists an 'evaluation gap' concerning microfinance-programmes: 'MFI field operations have far surpassed the research capacity to analyze them, so excitement about the use of microfinance for poverty alleviation is not backed up with sound facts derived from rigorous research. Given the current state of knowledge, it is difficult to allocate confidently public resources to microfinance development' (Zeller and Meyer 2003). Moreover, even if microfinance was the object of a significant number of impact evaluations in the last decade, it is one of the first times that a randomised controlled trial has been used. Finally, it is particularly interesting to determine the microcredit impact in rural areas because reaching this category of population, which is amongst the poorest, is a challenge for many microfinance institutions.

Al Amana, which was created in 1995, is the largest microfinance institution in Morocco, serving 250,000 clients. Until 2006 its clientele was mainly from urban or peri-urban areas (accounting for 83 per cent of the clients), but now Al Amana's strategy is to serve the rural areas at a significant level. After opening approximately 100 branches in the easily accessible hinterland in 2004 and 2005, Al Amana decided to expand into the enclosed rural regions. In order to rigorously measure the impact of microcredit distribution in this challenging new context, Al Amana management asked AFD for financial support to conduct this study. The institution had already identified the Poverty Action Lab as the research team that would be in charge of the evaluation, in partnership with the newly created Paris School of Economics.

The objective of the research programme is to analyse the economic and social impacts of microcredit in enclosed rural regions in Morocco, using an experimental method (Paris School of Economics 2006). The randomisation of the treatment assignment, with a group which will be exposed to microcredit from the beginning and another group later, will give clear, transparent and rigorous estimates of the impacts. The roll-out of Al Amana makes a perfect context for this type of method.

The evaluation concerns 80 of the 160 branches that Al Amana has planned to open between 2006 and 2008. The principle of the study is to identify two small zones in the area covered by a branch, one zone being served immediately and the other one being served one year later, as initially decided. In this scenario, three surveys are to be conducted: a baseline, an intermediary survey after one year and a final survey after two years. The last survey is going to allow measurement of the effects of two years of credit distribution compared to one year in the control group, therefore giving a differential analysis of short- and medium-term effects of the treatment on the populations.

The *modus operandum* of the establishment of the two groups was specified thanks to a feasibility test first carried out on nine sites scattered throughout Morocco. It was not possible to simply draw villages from a list because there is great diversity in the rural settings, notably linked with the type of land tenure, crops, landscape, and climate. Moreover, it was important that the *douars* (Moroccan villages) in the control group be distant enough from the place where the microcredit branch was based, and from any other source of credit, so that contamination would be avoided as much as possible. The villages in the treatment group therefore had to be chosen in the same type of situation.

The feasibility survey helped to define a matching method for choosing a pair of villages (treatment and control) with the same characteristics on the basis of variables such as accessibility, population, main crops cultivated, etc. The random selection is eventually made for each branch on a pair of similar villages, randomly assigning one to the treatment group and the other to the control.

Feasibility again played a fundamental part in defining the procedures for the experiment, which was to construct a model predicting the villagers' propensity to take credit. In order to limit the number of interviews needed to achieve statistical power, it was crucial to be able to select as many future borrowers as possible amongst the households interviewed during the baseline survey. The feasibility study was conducted by interviewing 2000 households in the nine pairs of villages, and following the distribution of credit over the next six months. This observation phase of the take-up provided the information necessary to construct the predicting model.

In the subsequent villages, this model permitted prediction of the 25 households with the highest propensity to take up with Al Amana on the basis of a short questionnaire (10 questions) applied to 100 families.

The data collected at the first nine sites showed that the evaluation process was progressing successfully. The randomisation had worked well and the original differences between the households in the treatment and the control groups were not significant. Moreover, the collaboration between Al Amana and the research team had been exemplary. Nevertheless, several technical problems arose and led to a change in methodology (Paris School of Economics 2007).

The surveys conducted during the first year revealed that take-up was lower than theoretically expected in these enclosed regions where Al Amana had no previous experience. In the first branches concerned by the feasibility, the borrowing rate was 21 per cent after 14 months (36 per cent amongst the households of the treatment group, thanks to the propensity model). In the 23 subsequent branches included in the study (accounting for 39 *douars* in the treatment group), this rate was only 7 per cent on average after 5 months, with great heterogeneity amongst the villages (this rate varying from 0 to 55 per cent).

The lack of borrowers in the treatment group interviewed is potentially a threat to the final possibility of measuring a limited impact with statistical significance. With a borrowing rate of 20 per cent, it would be impossible to detect a change in consumption smaller than 21 per cent. In order to address this problem for the purposes of the evaluation, several steps were initially taken to sensitise the villagers in the treatment *douars*: the number of information meetings was increased, the quota of credit reserved for women was opened to any borrower, and incentives were given to Al Amana staff to serve these remote villages.

As it happened, the protocol had to be revised more profoundly. These difficulties showed that the one-year exclusion delay would not be enough to obtain a significant

difference between the treatment and control groups. It was then decided to extend the exclusion of the control group for a two-year period. This decision is not without consequences. For the Al Amana local staff, it means explaining to the population of the control village that the credit is delayed. It also means that Al Amana is serving fewer clients, therefore leading to a shortfall for the institution. Since the mid-term survey was cancelled, its budget was reallocated to include 20 more villages in the survey in order to have more opportunity to reach statistical significance in the end.

These substantial adjustments have been possible thanks to a remarkable partnership built between Al Amana, the Paris School of Economics, and AFD to overcome the difficulties and achieve the evaluation. The meetings between all the parties are regular and help prevent any misunderstanding. The research protocol is highly transparent for all stakeholders, and the results will be available in 2010.

2.4 A Randomised controlled trial of micro-health insurance in Cambodia

Health insurance is one of the most important policy issues facing the developing world today. France recently pledged to invest more in Social Protection in developing countries. AFD is relatively new to financing health programmes, and sustains only two micro-health insurance programmes: one in Cambodia and one in Laos.

In order to know more about this before scaling up, AFD decided to launch an impact evaluation of the SKY Health Insurance Programme that it finances in Cambodia. Started in 1998 by Groupe d'échange et de recherche technologiques (GRET),¹ SKY offers households, for a fixed monthly premium, free and unlimited primary and emergency care at health centres, as well as a number of other health services. One of SKY's primary goals is to enable families to cover health costs without risking impoverishment.

In 2005, AFD signed a memorandum of understanding for the execution of project evaluations with Scientific Evaluation and Global Action (SEGA) of University of California at Berkeley and University of California, San Francisco. After a first methodological proposal was written by SEGA about the impact evaluation of the SKY project, an identification mission took place at the end of 2006 to more precisely define the possible methods and the scope of the evaluation, and to start fostering a buy-in of the future conclusions by policymakers in Cambodia.

The core method of the impact evaluation of the SKY micro-health insurance project is a randomised controlled trial (Levine *et al.* 2007). It will be implemented as SKY rolls out to Takeo province, currently scheduled to begin in approximately June 2008. For the preferred study design, the central methodological tool is the randomisation of coupons for premium reductions to vary the likelihood of insurance take-up among households within a village and isolate the impact of health insurance on the outcomes of interest.

Following the initial village meeting, when the coupons are randomly distributed, the baseline survey will be administered to a random subset of households, stratified by coupon value. From the baseline survey data and SKY's records of which households opted to take up insurance, it will be possible to answer the questions regarding

participation in the insurance programme. For example, it will reveal which household characteristics predict take-up. Furthermore, since the premium is randomly assigned, it will be possible to assess how premiums affect the baseline characteristics of insured vs. non-insured households.

Twelve months after each village meeting, follow-up surveys of all households originally interviewed will be carried out. The follow-up and the baseline data will give information on how SKY affects health-seeking behaviour and healthcare utilisation, as well as on how health insurance affects economic outcomes, such as changes in out-of-pocket expenditures. A second follow-up a year later will repeat most of the same topics, again emphasising changes in health outcomes and expenditures.

Since longer-term effects of insurance are also very interesting, high dropout rates among large coupon winners is a concern. If pilot tests show that most people who win a coupon in the first period renew their insurance for an additional 6 months, the above design will suffice for effects over at least the first 12 months.

But, as in any project, the evaluation might not go entirely as planned. The main threat is the number of households that will not continue their membership with SKY after their initial six-month period. Currently, the dropout rate for SKY after the initial six-month period is approximately 17 per cent (based on past SKY records). If the dropout is substantially higher than this for purchasers who received high-value coupons, there may quickly be little difference between insured status of the initial treatment (high-value coupon) and control (low-value coupon) groups. It may then be necessary to administer a non-experimental 'matching' method to gauge SKY's impacts.

In addition to household surveys, the research team will also administer a qualitative evaluation of the SKY programme. This analysis will examine the impacts SKY has on the healthcare system, including public health facility revenue, changes in supply of drugs and medical equipment, and changes in health-worker income and work patterns.

In the end, randomisation will allow the researchers to credibly estimate the causal effects of health insurance, as distinct from all other characteristics that vary across insured and non-insured households. A pre-intervention baseline survey of approximately 3,000 households with over 15,000 individuals and follow-up surveys of the same households will be conducted over the four-year experimental period. The survey will cover the multiple areas that the programme aims to influence: health status, health-seeking behaviour, asset vulnerability, investment and saving decisions, and risk management. Drawing upon the randomised research design, it will be possible to compare the changes in outcomes over time across insured and non-insured households to estimate the causal effect of health insurance.

A key feature of this impact evaluation is a series of partnerships both within Cambodia and globally that ensure the evaluation will match well with the needs of the programme, its funders, such as AFD, and other stakeholders with an interest in healthcare delivery to the poor.

Throughout the development of the methodological proposal, SEGA has been working closely with staff from AFD and GRET as well as Cambodia-based research partners at Domrei Research and Consulting. GRET's input in particular has been essential for all aspects of the proposal, including determining the feasibility and relevance of research designs. These relationships are being developed as the research design is structured and implemented and as the survey instruments are installed.

In addition, SEGA have been awarded a USAID-funded grant, which will allow them to develop capacity-building activities for local researchers and practitioners in Cambodia. These activities involve partnering with two researchers from the Royal University of Phnom Penh (RUPP). Throughout the evaluation period, SEGA will offer training in programme evaluation design and methodology to RUPP students, with the aim of enabling future impact evaluations to be run locally. Moreover, part of the budget allocated by AFD includes presentations in Paris to disseminate SEGA's methodologies (including research design and econometric techniques) and findings.

As verified during the first mission to Cambodia, an evaluation of the SKY programme directly supports the goals of the Cambodian Ministry of Health. In particular, knowledge about SKY's effectiveness will help the Ministry to structure its reform of the Cambodian healthcare system – ongoing since 1999. During this mission, the evaluation team and AFD met with Ministry of Health officials to collect information on questions they find particularly important in the SKY evaluation. Throughout the evaluation, the input of Ministry of Health officials will be sought regarding the study design and they will be regularly informed on results, which will become available in 2010.

3 Intermediary lessons

3.1 Institutional learning

From the outset, AFD has progressively learned and changed its approach to impact evaluations. The main evolution has been the adoption of methods necessitating the construction of a baseline, which has meant convincing the various stakeholders and identifying the academic partners long before a project is launched. The preparation of such research is obviously time-consuming: it means having policymakers, researchers, data collectors, development operators, and donors meet and agree on the principles and details of the exercise. In the Moroccan and Cambodian cases, it meant dedicating one and a half years to this preparation phase.

Progress has also been gradually achieved in the incorporation of local partners. It is very clear that impact evaluations are a very demanding exercise and need the full involvement of the project team being assessed. The contribution of the project team is essential in contextualising the questionnaires and in adjusting the sample size in line with the expected take-up. The project team's involvement is all the more crucial during the impact evaluation's implementation, in order to correctly apply the design, particularly when it comes to preventing contamination of the control group. It also often means that the

project and the various stakeholders must be ready to change the intervention protocol to allow for randomisation or for building a good comparison group.

Yet, the strategic objective of implementing impact evaluations to contribute to policymaking means more than involving the project team. The rigorous information might not make a difference if it is not taken into account by national policymakers. Particular efforts were made in the Guinean and the Cambodian evaluations to promote a political buy-in from start to finish.

Based on current experiences, impact evaluations appear as much a research product as an evaluation product. In their expected feedback they are close to evaluation but the nature of investigation is clearly research. They provoked new opportunities for AFD to collaborate with high-calibre academic partners in the analysis of its operations. A strong in-house involvement in every exercise, as much from the project managers as from the Evaluation Unit, develops an evaluation culture that forms a necessary basis for the launch of more evaluations of scientific quality in the future.

3.2 Impact evaluation for a bilateral donor

AFD's experience in impact evaluation (IE) might appear minimal. Two exercises will be in progress in 2008 (Cambodia and Morocco) concerning only two AFD-supported projects amongst about 500 ongoing operations. However, these works will consume 25 per cent of the Evaluation Unit's budget for outsourced evaluations and between 10 and 15 per cent of its human resources. This imbalance deserves attention.

For a bilateral donor, the first question about impact evaluation is to wonder if it is worth engaging in. The second question would be about whether to pilot it internally or outsource it. If the internal option is preferred, then the third question would be where to locate the IE activity: an evaluation unit, in a knowledge department or in a policy department?

As explained above, AFD has tried to address these questions by launching a few impact evaluations piloted by its evaluation unit. The last question on IE location was not a problem for the AFD due to the positioning of the Evaluation Unit inside the Research Department.

With the benefit of hindsight, an internal involvement in piloting IE seems a good way to fuel the debate on impact measurement inside the institution. It seems barely possible to develop results-based management without experiencing what a rigorous impact assessment means.

Moreover, direct implication of AFD in the management of IE appears crucial to fully understanding the dos and don'ts and objective difficulties and challenges inherent in rigorous impact evaluations, as aforementioned examples show. Internalising IE or outsourcing to a pool dedicated to this could be a false alternative. Appropriation of IE is a prerequisite to having fruitful participation in international networks, sharing not only results but also processes and methods.

3.3 Perspectives

Our analysis of the Randomised Controlled Trials (RCTs) launched so far is that, although they are very interesting, they present several obstacles when applied to the type of projects that AFD finances. A recurrent difficulty in applying quantitative methodology with a baseline to AFD projects is the self-selection of beneficiaries and the slow take-up where new specific services, like microcredit or micro-insurance, are proposed for the first time. Moreover, there is a contradiction between maintaining a control group without contamination and measuring the impacts of a new project, which can take time to materialise, on individuals as well as on a population as a whole.

The results of the recently launched RCTs are expected in only a few years' time. Each of them required 18 months of identification before starting. Their cost, their demanding preparation and implementation, as much as the aforementioned objective threats to the evaluation design and the uncertainties of their impact on policies, encourage AFD to wait for the results before engaging in another exercise of this type.

One of AFD's concerns for future work would also be choosing methods that permit answering the maximum number of relevant questions at a policy level. Mixed methods were or are going to be used in the Guinean and the Cambodian evaluation for this purpose: information on process and on impacts on key stakeholders other than the direct clients of the project will be collected and analysed. But it remains difficult to identify researchers who are inclined to use a combination of methods as rigorous in impact evaluation already means dedicating much effort to the fine tuning of the principal method used. It is a responsibility of the project and of AFD to be sure that the need to answer the maximum number of relevant questions prevails in method-oriented research.

In fact, the Evaluation Unit at AFD now considers it a challenge to contribute to methodological innovation for rigorous impact evaluations, using a counterfactual and quantifying the net change in outcomes but not necessarily using experimental or quasi-experimental techniques. A variety of rigorous impact evaluation methods is necessary to answer the array of questions raised by development interventions while at the same time addressing attribution. As it was demonstrated with the Guinean Impact Evaluation, the AFD Evaluation Unit believes that qualitative researchers can rigorously address the impact evaluation challenges.

In order to face the growing in-house demands for impact assessments, AFD intends to adopt a pragmatic approach, promoting more quality evaluations in general and, when possible, integrating the attribution question. Few project managers are ready to finance baseline and specific data collection, and fewer of them are presently willing to engage in a long process and to dedicate a significant budget for a research programme on the impact of their operations. More information is clearly needed about the real effects of impact evaluations before developing them widely.

ⁱ Research and Technological Exchange Group.

REFERENCES

AFD (2007) '2007–2011 Strategic Plan of the Agence Française de Développement', Paris

Bamberger, M., Rugh, J., Mabry, L. (2006) *Real World Evaluation: Working under Budget, Time, Data and Political Constraints*, Thousand Oaks, CA: Sage

CGD (2006) *When Will We Ever Learn? Improving Lives Through Impact Evaluation*, Washington DC: Center for Global Development

Delarue, J. (2007) 'Mise au point d'une méthode d'évaluation systémique de l'impact des projets de développement agricole sur le revenu des producteurs. Etude de cas en région kpèlè (République de Guinée)' PhD thesis, Agroparistech, Paris

Gittinger, J.P. (1982) *Economic Analysis of Agricultural Projects*, EDI Series in Economic Development, Baltimore: Johns Hopkins University Press

Gubert, J. and Roubaud, F. (2005) *Analyser l'impact d'un projet de micro-finance : l'exemple d'ADÉFI à Madagascar*, Notes et Documents, numéro 19, Paris: AFD

Levine, D.I., Polimeni, R., Arunachalam, R. (2007) *A randomized controlled trial of micro-health insurance in Cambodia. A preliminary proposal to AFD*, SEGA, Berkeley: UC Berkeley

Paris School of Economics (2007) 'Evaluation de l'impact du microcrédit en milieu rural' Note préparée pour le Comité de pilotage du 1^{er} août 2007, Paris: Paris School of Economics

Paris School of Economics (2006) 'Evaluation de l'impact d'un programme de microcrédit en milieu rural' Research Project, Paris: Paris School of Economics

Zeller M. and Meyer, R.L. (dir. pub.) (2003) *The triangle of microfinance: financial sustainability, outreach and impact*, Baltimore: Johns Hopkins University Press