Impact Evaluations: a tool for accountability?
Lessons from experience at AFD

By Jean-David Naudet, Jocelyne Delarue et Tanguy Bernard

“If you have a hammer, everything looks like a nail”
Abraham Maslow, the Psychology of Science, 1966

Abstract

This paper relates the Agence Française de Développement's experience with respect to impact evaluations. Our purpose is to assess the extent to which such studies, when designed before an actual program implementation can provide the type of summative evidence that donors still seek for when promoting such studies. Specifically, we rely on three large scale randomized control trials, and scrutinize their capacity to answer questions related to the underlying program’s impact ‘on whom’, ‘on what’ and ‘of what’. We conclude that experimental studies should be promoted to answer the type of ‘tunnel’ questions characterized by a limited number of well-specified homogeneous inputs, a tried and tested process, a short and external events-proof causal chain, a rapid and stable appropriation by beneficiaries, a large and stable participation, and a set of measurable outcomes in the short run. While a number of such issues exist and are very much worth studying experimentally to inform future development policies, few development interventions satisfy them, and summative use is thus limited.

1. Introduction

Since the mid-1990s, donor agencies have been increasingly concerned with the demonstration of their capacities to improve the lives of the beneficiaries of their interventions. Epitomized in the Paris Declaration for Aid Effectiveness, donors have notably pledged for a shift of focus towards development results and their measure (OECD 2005). And today, despite notable controversies on attribution, one finds assessments of numbers children sent to school, of farmers trained, of malaria-related death avoided, and eventually of households lifted out of poverty, on most donor agencies websites.

Yet, by the early 2000, it was increasingly recognized that academic literature on growth and its relationship with aid had mostly failed to establish the type of causal relationships needed to assess development policies (e.g. Easterly, 2003). And at the micro-level, evaluations of particular interventions also lacked the necessary ‘robustness’ to deal with all sorts of biases when assessing their impacts on beneficiaries (e.g. Duflo & Kremer, 2003; Banerjee, 2006). Thus, donors’ debates over policies or projects were deemed by some as having the characteristics of ‘ignorant armies clashing by night’ (Pritchett, 2002): heated debates occurred without any firm evidence to argue for or against the likely final impact of given interventions, and donors responded less to evidence than to political fashion (Deaton, 2007).
Under the leadership of several academics, impact evaluations (IE) akin to those used in the medical field, based on comparisons of adequately defined treatment and control groups, have since been proposed as a means to reliably estimate causal relationships between interventions and their outcomes. With such robust and cumulative evidence, their promise is to help foster the missing links between academic research and development practices, to promote trial and errors process in policy building with experimentations preceding scaling-ups, and to finally deliver knowledge on ‘what works’ in development policies (CGD 2006). And overall, the donor community has largely followed this call, collectively contributing to dedicated international funds, and using IE to “strengthen (their) internal monitoring and evaluation systems,” as recommended by a report from the Center for Global Development (2006).

Yet, it is fair to say that in a context of strong demand for results, one of the main motivations of development agencies has mainly been to use impact evaluations as a means to demonstrate their effectiveness and only secondarily to use them as a learning tool. For instance, a recent ODI survey on impact evaluation production and use reports that “the available evidence does seem to point towards experimental IEs being commissioned in order to fulfil accountability purposes” (Jones and al 2009). And in fact, most reference books continue to highlight the special power of IEs to contribute to better accountability for development institutions (Khandker et al 2010, Gertler et al. 2010). Accountability, in turn, suggests that IE provide summative measures of impacts as for instance defined by the donor’s community itself: “All positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended” (OECD/DAC 2002).

While recognizing the significant improvements that experimental methods have brought to empirical studies in the field of economic development, this paper suggests that donors’ use of IE for direct accountability purposes may be limited. In particular, we posit that treatment-control methods are best suited to address ‘tunnel’-like issues characterized by a clearly defined and stable ‘treatment’, a rather short and event-proof causal chains, and a large share of targeted individuals that are effectively affected by the intervention. In fact, this ensures that IE effectively delivers responses to three questions regarding the intervention’s impacts: impact of what, impact on what, and impact on whom. Yet, most development interventions do not satisfy such pre-requisites, limiting the use of IE for donors’ accountability needs. In essence, while impact evaluations are well-suited to help understand development processes and test the mechanics linking an intervention to a given outcome, they are less often appropriate to donor’s needs for measures of their impacts in the field.¹

The argument is illustrated with the Agence Française de Développement’s (AFD) own experience with impact evaluations. Following the Center for Global Development’s call for increase IE, AFD has piloted several such IE on its operations over the recent year, among which

¹ Our focus here is on the use of impact evaluations from a donor’s perspective. We do not seek to address the subject of impact evaluations and among them RCTs from a methodological standpoint, which is already the subject of a large body of literature (see for instance a symposium edited by the Economic and Political weekly edited by Ravi Kanbur (2005), a Boston Review book edited by Abhijit Banerjee (2007), a 2008 conference held at the Brookings Institutions in Washington D.C., or the World Bank blog on impact evaluations (http://blogs.worldbank.org/impactevaluations/) for glimpses of on-going debates).
two large scale randomized control trials related to micro-credit in Morocco, and health insurance in Cambodia. While AFD’s main purpose was to learn more about this new evaluation methodology, it also hoped to provide robust evidence of the impact of its interventions. Using the world-renown Progresa impact evaluation as a benchmark, we assess the extent to which AFD-supported studies were able to provide summative assessments of the interventions’ impact on their targeted individuals.

The rest of the paper is organized as follows. Section 2 briefly describes the three studies underlying the argument – namely Progresa, Al Amana and Sky. Section 3 discusses the extent to which the studies were capable of answering questions such as: the impact of what, on what, and on whom. Section 4 identifies common features and differences of the three analysed examples with respect to the IE design and characterizes the types of ‘tunnel-like’ issues that are most appropriate to experimental studies. We conclude with a call for donors’ continuous support for experimental studies, but for reasons more related to learning about policy mechanisms and policy building through experimentation, rather than fulfilling accountability needs.

2. Three impact evaluation cases for study

Over the past six years, the Agence Francaise de Developpement has engaged in the financing and piloting of a few impact evaluations (IEs). The Agency’s main objectives were to provide results of the projects under investigation for accountability purpose, but also to assess the potential of these new evaluation tools toward the production of relevant knowledge for the institution’s various needs. More specifically, AFD’s work was geared at a better understanding and appropriation of the results generated, favouring their utilization and dissemination.

This paper focuses on two large-scale randomized control trials (RCT) performed at AFD. In fact, it is these approaches that have mostly been promoted throughout the impact evaluation movement – they now represent the vast majority of studies being undertaken. Focusing on these two studies eases comparisons with the impact evaluation reference, Progresa, which is briefly described below.

2.1. Progresa conditional cash transfers

Progresa (now Oportunidades) is a Conditional Cash Transfer (CCT) program, which has been implemented in Mexico since August 1997. A key feature of the program consists in providing cash transfers to families as an incentive for them to adopt human capital enhancement behaviour for their children. This means sending children to school and bringing them to health centers. The condition attached to the transfer is thought to be a crucial component, as it is expected that simple income effects of transfers do not necessarily translate into more schooling for children. Importantly, transfers were given to women, on the grounds that they tend to use resources for better nutrition, and it was also hoped that this would help emancipate them from their husbands’ authority. Finally, a number of supply side actions were also undertaken to ensure that enough schools and clinics were in place to fulfil the enhanced demand generated by the program.
As such, Progresa’s program theory was very simple and clear: address poverty in the short run, while tackling poverty in the long run via investments in human capital. Most results were expected in the short run and were easily quantifiable, be it school participation or preventive health behaviour. Other expected outcomes, such as empowerment of women, were less direct and expected in the longer run and were measured through proxy indicators.

The Progresa impact evaluation started in 1998 and relied on a large-scale randomized experiment: a subset of the program’s targeted communities was phased-in following a random order. The impact evaluation was one of the first such large RCTs on development intervention, relying on a sample of 24,000 households, and was therefore highly publicized. Nearly all households that were offered the CCT did take it. Within 18 months, results showed that the program had a number of positive impacts on health outcomes and school attendance (see Skoufias, 2000, for a review).

Progresa is a key milestone for the impact evaluation movement, for both the rigour of the methodology it rested on and the effects of the evaluation itself. Based on the results obtained, the program was continued and expanded to urban areas despite a change of government. Further, it was the basis for the promotion of other CCT programs throughout the world, most of which have also included an impact evaluation – providing the necessary basis for the ‘robust’ kind of meta-analyses targeted by the impact evaluation movement (Fizbein & Schady, 2009).

2.2. Al Amana rural microfinance

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

The design of the study was based on pairs of nearby and “similar” villages, from which one was randomly selected for immediate access to microcredit, while the other village was used as a control village for two years. A total of 88 pairs of villages were selected all over Morocco in order to provide representativity of results on a national scale. About 6,000 households were to be surveyed on three occasions (before Al Amana begins its activities, a year later and two years later).

---

2 In 2009, 28 such programs were implemented in developing countries, from three ten years earlier.
3 Other outcomes such as women’s empowerment or kid’s education were also planned as secondary questions.
4 The study design largely followed that of another impact study of microfinance in Hyderabad, India. Other such impact studies at the time were in preparation, including the one in Hyderabad (see Banerjee, Duflo, Glennerster, & Kinnan, 2010) and another one in the Philippines (see Karlan & Zinman, 2010).
A number of difficulties and changes of orientation arose throughout the project and the research process. The most important aspect concerned the take-up rate. Original expectations were mostly drawn from Al Amana’s own experience in urban areas and rural participation was expected to be similar. With this setting, a conservative estimate of 60% participation rate in the population in general would only be able to detect a 20% change in the final consumption level (Crépon, et al., 2006). In order to enable detection of smaller effects, the sampling scheme selected those 25 households per village that had the highest probability of becoming borrowers of the newly installed microcredit scheme, based on a propensity score parameterized using the data collected in the feasibility study (Crepon et al., 2007). Rapidly, however, the take-up rate was found to be much lower than expected. Several actions were therefore taken to try and enhance it: the intermediary survey was cancelled to enable take-up rates to rise, and some of the product’s features were changed (removal of quotas for women, modification of repayment schedule, enhanced information of population).

Results of the study show that, given sample size and take-up rates, no impact could be found, on poverty, consumption, activity diversification, or shock absorption, even if some significant effects on production and wages were identified – for those households above the median poverty level in particular (see Crépon et al., 2010a).

2.3. Sky health insurance

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

The SKY impact evaluation uses a randomized control trial to examine the causal effect of the proposed insurance on households’ economic and health outcomes and healthcare utilization decisions, as well as to understand who does and who does not choose to purchase insurance – in particular, to uncover aspects of adverse selection that are omnipresent in the insurance literature. To this end, the study relies on the random allocation of discount vouchers toward the purchase of insurance for six months. Such a setting allows for the comparison of those households that did contract insurance thanks to the voucher with similar households that did not contract insurance but would have, had they been given a voucher.

Here also, a number of difficulties arose. Dropout rates were significant after expiration of the discounted period, which led to an extension of the voucher scheme. In addition, the study aimed at assessing how insurance would reduce indebtedness at the time of severe health shock, an event too rare to detect statistically. Finally, the collected information regarding health services

---

5 That is about 17.6% on average after 24 months of the program being available to eligible individuals according to Al Amana’s administrative data, 10.6% according to the survey data, and 13.6% among the 25 households per village with the largest probability to borrow (Crépon et al., 2010b).
consumption was too parsimonious to derive meaningful measures of changes in behaviours, and little improvement in health outcomes was identified. The study does however identify some economic improvement in the life of those households where one of the member is sick. Finally, the evaluation uncovers some signs of adverse selection in participation to the insurance scheme.

3. Towards accountability? Summative impact evaluations

We use the three studies described above to illustrate IE’s potential to provide summative assessments of a programs’ impact. In a nutshell, summative evaluations entails that the study is capable of assessing the program’s overall impact under (close to) normal conditions. Specifically, we ask three different questions and use the three studies described above to provide some elements of an answer: On whom is the impact measured? Of what is the impact measured? And on what is the impact measured?

3.1. On whom is the impact measured?

Impact evaluations rely on sampled units, which may or may not be representative of the underlying population of beneficiaries. A central feature here is the participation rate. In fact, and quite logically, only those units that participate in the program’s treatment group, and those from the control group who would have participated had they been offered the program, help estimate the effect of the intervention on its beneficiaries. However, in cases when participation is not mandatory, and thus the participation rate is not 100%, it is not possible ex-ante to distinguish with certainty those units that will participate from those that will not. Because of this, the best way to draw a sample that is nevertheless representative of all participants is to randomly select units from the pool of all targeted beneficiaries.

If the participation rate is high, most of the randomly drawn sample will be useful for the estimation. This is the case of Progresa, which has a participation rate of more than 98%. If the participation rate is limited, however, only a fraction of the sample will help estimate the impact, which in turn could limit its statistical capacity to detect an impact of reasonable importance. It is sometimes argued that one could nevertheless assess the impact of the intervention on the targeted population – instead of the effective participants – which would take into account eventual externalities from participants to non-participants within the targeted population. And in fact, this estimator – the intention-to-treat estimator – may be more relevant from a donor’s perspective as it produces an answer to questions such as “to what extent has well-being in this region been changed as a whole thanks to the intervention” (Ravallion, 2008). Yet, unless externality effects are strong and widely disseminated among the targeted population, it is unlikely that the IE will be able to capture these effects in the case of low participation rates.

To account for this participation issue, a very purposive sample selection scheme was implemented in the Al Amana study (cf Section 2), whereby only those individuals with the highest borrowing probabilities would be included in the sample. In cases where the predicted probabilities included the vast majority of effective borrowers within the sample, the obtained impact estimates would in effect be representative of the total population of borrowers. Yet, the models used for the purposive sampling most often have positive but limited predictive
capacities and are only capable of marginally increasing the rate of participants in the sample. In the case of Al Amana, the borrowing rate in the targeted population was of 11%, compared with 13% in the sampled population. Overall, the study’s population is clearly not representative of either the population of targeted households or of the actual participants. Further, the complexity of the model used for prediction limits the study’s capacity to replace sampled individuals within the general targeted population. As a consequence, the evaluation informs us on the (non)existence of an impact and its magnitude, but for a very specific and purposely built sample of the population targeted. In effect, it doesn’t provide a measurement of the program’s impact on a sample representative of the targeted beneficiaries, which is the key information needed for accountability purpose.

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

Here again, however, it is worth questioning to what extent the population on which the treatment is estimated is representative of the population on which the program is normally implemented. And in fact, by design, impact is assessed in this case on a population that would not have chosen to participate had the insurance been priced at normal rates. In fact, the ‘always-takers’, those households that would participate under normal conditions, do not contribute here to the impact measure – this is akin to what is sometimes referred to as ‘randomization bias’ (e.g. Ravallion, 2008). Overall, it is quite likely that the impact obtained is different for those who are willing to purchase the insurance at a higher price than those who only join when it is discounted. In consequence, the magnitude of aggregate causal effects on all beneficiaries of the insurance remains ignored at the end of the evaluation. Here again, a key information is thus missing if one is to use the study for accountability needs.

Overall, while in the case of Progresa, the study sample was likely representative of both the targeted and the actual beneficiaries, this is less the case for Al Amana and Sky. In these studies, responses to weak participation rates and the need to randomize at individual levels led to purposive or partial samples that are no longer representative of populations that are relevant for donors’ accountability needs (be it the targeted population or the population of effective participants).

3.2. The impact of what is measured?

---

6 If one wanted to measure the impact of those who participate under normal conditions, the best thing to do would be to provide negative incentives to some, and normal conditions for others (such as higher prices than normal). In such case, the pool of compliers is likely representative of those who participate under normal conditions. Obviously however, it would raise ethical issues, to make people pay more than in normal conditions for services that we know are potentially good, such as health insurance.
We now turn to the project that is being evaluated in order to investigate how this can be used for accountability purposes. In fact, we shall see that the impact that is evaluated may not be as easily interpretable as of the impact of the project under evaluation.

We first ask whether the project implemented can be interpreted as if it were implemented under “normal” conditions. In the case of Progresa, the IE only examined a subset of the communities targeted by the Program, while Progresa was at the same time being rolled out throughout the country under the same design. Overall, one can feel rather confident that the IE produced results very much representative of the program as implemented under normal conditions.

The situation was different in the case of Al Amana, where the program was very much implemented without its adaptation to new clientele being fully thought off. In fact, it was the first time that Al Amana entered the targeted remote rural areas, and a number of issues of the program itself remained to be addressed. Accordingly, Al Amana had decided to first implement the scheme the way it was implemented in urban areas and to later modify it on the basis of the learning generated through experience.

For instance, the repayment calendar was initially designed to mimic the way Al Amana provided microcredit in urban areas, without taking into consideration the agricultural calendar. However, over the course of the evaluation, some loans were later made available without repayments for the first few months. Further, quotas for women were initially implemented so as to ensure that women were encouraged to participate. This was later abandoned through the course of the study because of the low take-up rate observed. Similarly, while it was initially planned that the credit would be provided through groups, this later changed to allow some individuals to borrow individually. Finally, despite the original call for a normal level of households’ information on the product, an observed weak take-up rate also led Al Amana to provide extra incentives to its field agents for the duration of the study.

In sum, Al Amana’s project was not stabilized at the time of the study and was very much considered an on-going learning effort by the association. Because the IE had to be implemented at the same time as the opening of new branches, it led to the study of the impact of a very variable product, that doesn’t make very much sense for Al Amana itself. As such, while the study’s aim is to determine the impact of access to Al Amana’s microcredit scheme in general, it must be qualified as the impact of an unstabilized, and thus not fully fledged, intervention.

In the case of Sky, however, the project had already been piloted for several years in other communities; thus, the intervention was now mature and stabilized. As such, the IE could effectively investigate the impact of the program as normally implemented. A caveat must nevertheless be added in that it is no longer the impact of insurance that is being assessed, but rather the impact of an almost free insurance product that was provided for the sake of the study. This may have consequences, as one may expect that beneficiaries’ reactions to this provision are different. In a sense, the question addressed by the evaluation is the effect of “being (almost) offered an insurance” and not the effect of “taking an insurance”. In terms of accountability, the distinction could make a difference.

\[7\] See Barham 2005 for an in-depth description of Progresa’s Rollout.
Overall, while the Progresa study evaluated the intervention as it was and as it was planned to be in the future, both Al Amana and Sky evaluated programs that are different from normal implementation conditions. This obviously limits the use of the results for summative evaluation purposes.

We now turn to the issue of the variable intensity of the program on different beneficiaries. By intensity, we mean the amount of transfer (Progresa), microcredit (Al Amana), or insurance (Sky) that different households have had access to. In the case of Progresa, intensity was given by design, in that households were entitled to a fixed amount of transfer, based on family demographics, and could not change that amount. In the case of Al Amana and Sky, however, households themselves could decide how much of the program they would ‘consume’, which results in important endogenous treatment intensities. In Al Amana, some households may have borrowed different amounts or may have borrowed several times, while others may have only borrowed a single time. Further, take-up was observed to be very progressive within communities, which means that some households used the credit early on while others had barely started to borrow at time of the second round survey. Sky offers a similar issue in that a lot of households dropped out of the insurance scheme over the course of the study, after the grace period was passed. As implemented, however, the study pulls together households that have been insured for two years with households that were only insured for six months.

In the cases of both Al Amana and Sky, these non-constant treatment intensities clearly entail limits to the interpretation of the studies’ impact results. Obviously, however, the analyses could try to differentiate the levels of impact according to different intensity of borrowing/insurance. Two limitations arise however: one relates to power in that the dataset that doesn’t allow to run impact estimates on subsamples. The other relates to endogeneity of intensity itself, that the experiment fails to account for.

A final issue relates to the novelty of the programs themselves. As mentioned before, Progresa happened to be quickly adopted by its potential beneficiaries. In contrast, microcredit in remote rural areas of Morocco was clearly not well understood by a large portion of the population. As a matter of fact, a qualitative study undertaken by Guerin et al. (2010) on a subsample of the study’s villages clearly showed that households had very varied perceptions in terms of the use of credit and repayment obligations (the latter, for instance, as a function of whether Al Amana was perceived as a government entity, such that credits were mostly understood as transfers). The same is true in the case of Sky, where households had to understand an entirely new concept whereby their health costs are covered by their insurance scheme. In both cases, it is likely that after a few years of experience, households will react very differently from their trial–and-error period of the first few months.

Overall, both the Al Amana and the Sky studies show significant challenges in their capacities to interpret the results. This again limits the use of these studies for summative purposes.

3.3. The impact on what is measured?

A final issue relates to the outcome that the program seeks to affect and how it can be measured within an impact evaluation. In the case of Progresa, clear and easily measurable indicators,
directly related to the MDGs, were established regarding participation in school and visits to health centers. Further, the impact on these outcomes was expected in the very short run, since the transfer would stop after three months if the conditions were not fulfilled. In other words, the causal chain was very short. The impact on Progresa’s other targets, such as empowerment of women, is more challenging to assess because of measurement difficulties, as well as the time it may take for impact to occur. In fact, the Progresa impact evaluation relied heavily on qualitative studies to assess these latter features (Skoufias, 2000).

The outcomes on which credit has an immediate impact, for instance agricultural production, are not as relevant to discuss contribution of the program to MDGs as in the case of Progresa. On the contrary, it is reasonable to expect that the impact of microcredit on such outcomes as poverty takes a relatively long time to materialize. In fact, the first credits start with small amounts and for the investments financed to be transformed into decreased poverty, a number of steps must be gone through. In addition, and as discussed previously, a learning curve is likely associated with the use of such a new product. Overall, it may take a relatively long time for credit impact to materialize into changes in poverty outcomes. Finding no statistically significant effects in the short run could therefore be misleading.

The issue is different in the case of Sky. Here, the main idea was to measure the impact of health insurance on health seeking behaviour, health-related outcomes as well as on debt after an accident or illness. It appeared that variations in the factors that are most related to MDGs (notably maternal and infant health) are very small and occurrences very rare, such that it was difficult to find statistically valid variations. And while the study finds impact on intermediary outcomes, the link between such outcomes and donor’s need for MDGs-framed results is not complete.

Thus, in both Al Amana and Sky cases, the summative use of IE results is limited by the limited time frame of the study, and the main outcome considered initially (poverty in one case, lower indebtedness in the other case) that are both expected to be statistically detectable after a significant amount of time. And in fact, IEs of the types described here, are most often weakly adapted to assess medium term impacts. The major constraint here is that it is difficult to keep a control group immune from the program for too long. And while some attempts are made to evaluate longer run effects of Progresa-like programs, they have to rely on variations in the number of months exposed to the program, which significantly affect statistical power.

4. Conditions for summative impact evaluations

What lessons can be drawn from these three examples with respect to the use of IEs for accountability needs? We use the differences in nature of the three above-described programs, and the questions effectively addressed in their respective IEs, to propose a rule of thumb

---

8 In fact, results from the Spandana study in Hyderabad suggest that the causal chain between access to microcredit and poverty is long, heterogeneous and complex (see Banerjee, Duflo, Glennerster, & Kinnan, 2010).

9 It has been argued that finding no short-run impact on poverty is itself an indication that programs with faster effects should be promoted instead. An in fact, it is true that programs such as social transfers may work faster. This however forgoes the recurring issues of sustainability of impacts themselves.
identifying the kind of programs and questions for which IEs can satisfy accountability requirements.

Let us first extract some key components on which these programs differ and are relevant to our discussion on impact evaluations. First, relating to “on whom the impact is measured” the three programs vary in beneficiaries’ propensity to participate and thus in their take-up rates. In the case of Progresa, a large part of the beneficiaries are expected to participate and actually nearly all targeted individuals chose to accept the payment and the conditions attached. In the case of Al Amana and Sky, however, only a fraction of the population was expected to do so, leading to very purposive sampling schemes that may affect the studies’ capacities to generate results that are representative of the underlying targeted population.

Second, relating to “the impact of what is measured”, the programs differ with respect to the novelty and complexity that they present to beneficiaries. Clearly, Progresa constituted an innovation for rural households in Mexico but results from the study indicate – a posteriori – that its mechanics (and in particular the conditions attached to the transfers) were fairly easily understood by the targeted beneficiaries and the program designed was not consequently changed over the course of the study. In contrast, the qualitative analysis undertaken in both Al Amana and Sky (cf Guérin et al., 2010; Portejoie et al., 2008; Ramage et al., 2010) revealed that the proposed products were rather new and complex. A slow learning process was thus to be expected on the demand side for the choice to participate, as well as for the use that beneficiaries would make of these products. This, in turn, may affect the type of impact that is measured.

Still with respect to this second point, homogeneity of treatment appears to be an important distinction between the programs. In the Progresa case, cash transfers are not homogeneous amongst beneficiary households, but the conditions attached to the transfers are the same for all. Homogeneity of the treatment could also be seen in terms of period, in that every beneficiary household within a wave of the program’s scaling up, is “treated” by CCT at the same precise period of time.

In contrast, heterogeneity of treatment characterizes Al Amana and Sky programs. We saw above that intensities of treatment were variable for beneficiary households according to the amount and the number of credit subscribed or the duration of insurance cover. Heterogeneity pertains also to the time where the treatment is taken. For both cases, credit and insurance could be taken by households at any time between baseline to the end line surveys. Importantly, these differences in intensity depend on the choice of the households themselves, not of the program. Finally, for Al Amana the conditions of the treatment and particularly the interest rates are different between households according to the period considered.

Third, relating to “on what impact is measured”, these programs differ in the expected length of time necessary for the intervention to reach its objectives. In fact, while Progresa, Al Amana, and Sky all target long-term poverty alleviation, short-term intermediary outcomes with a direct link to poverty (for instance through Millenium Development Goals) are more or less easily identifiable. In fact, the impact of Progresa on such outcomes as school participation or visits to health centers can be observed within just a few months (households would otherwise immediately lose access to the transfer). In contrast, the impact of microcredit on poverty may
take a longer time to effectively reveal itself, and the relationship with short-term impacts can be rather tricky (e.g. Banerjee, Duflo, Glennerster, & Kinnan, 2010). As for the case of Sky, impact is only to be measurable if significant health shocks occur, which can only be statistically detected if a rather large number is observed. Here also, measurable impacts are therefore more likely to be identified in the medium- to long-run.

Table 1. Selected differences between Progresa, Al Amana and Sky.

<table>
<thead>
<tr>
<th></th>
<th>Participation rates</th>
<th>Learning curve involved</th>
<th>Homogeneity of treatment</th>
<th>Length of time for impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Progresa</td>
<td>Close to 100%*</td>
<td>Limited</td>
<td>Strong</td>
<td>Short term</td>
</tr>
<tr>
<td>Al Amana</td>
<td>13.6%**</td>
<td>Important</td>
<td>Weak</td>
<td>Medium/long run</td>
</tr>
<tr>
<td>Sky</td>
<td>27% ***</td>
<td>Important</td>
<td>Medium</td>
<td>Medium/long run</td>
</tr>
</tbody>
</table>

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

Table 1 summarizes the features that are determinant for the capacity of IE to produce summative assessments and to meet donors’ accountability requirement. In particular, it highlights some of the reasons why the Progresa IE satisfied accountability needs when Al Amana and Sky evaluations were less able to.

Two precisions are worth mentioning at this stage. First, most of these constraints essentially apply to those impact evaluations that are planned before the program is effectively implemented. In fact, in such cases, and particularly in the case of innovative approaches, take-up rates are mostly unknown ex-ante, and the program’s design is not yet stabilized. Such constraints are also particularly relevant in the case of IEs which affect the intervention in itself, either through the choice of the beneficiaries/non-beneficiaries, or through an influence on the content of the program itself – such as the pricing policy of the service provided for instance. Randomized Control Trials such as the ones described above, while arguably the most statistically robusts, are also the most typical examples of these types of studies that are limited in their capacity to provide summative assessment of a program’s impact.

Second, it would be a mistake to draw from this table the idea that Progresa is a simpler program that the two other examples. Progresa is a large and complex program with several sub-components on both demand and supply side of health and education. What is perhaps simpler, or at least more focused, is the main policy question raised by the program. In fact, the Progresa IE has been mainly concerned with the evaluation of effects of the conditions linked to the transfer. In this sense, it is the evaluative question, and not the program in itself, that appears more manageable in the case of Progresa than in the cases of Al Amana and Sky.

Overall, the above discussion suggests that only a subset of development programs are good candidates for summative impact evaluations. A period of observation coherent with the logical chain, a limited number of well-specified homogeneous inputs, a tried and tested process, a short and proof causal chain, a rapid and stable appropriation, a large and stable participation, and a set of measurable outcomes in the short run and/or impacts covering the main aspects of the program, or at least the main aspects of evaluative questions, are conditions favourable to a
summative use of RCTs. A useful analogy to describe such programs is that of a tunnel, that has a delimited beginning and end, where one can easily define what enters in as an input and what is expected to come out as an output, where it is close to impossible to drop-out in the middle of the way, and finally where the path is both short and predictable in that it is immune from external influences. It will not come as a surprise that the characteristics of tunnel programs match nicely with features of medical experiment that are at the source of the impact evaluation movement.

5. Conclusion

It is a well known feature that not every program is adapted to impact evaluation, and particularly to RCTs. For instance, such methods only apply to programs where a large number of treatment units can be compared to a large number of control units. IE are therefore suited for micro-level types of interventions, such as those where beneficiaries are individuals, households, classrooms, or local communities. Thus, a share of development programs is not suited for IE.

The above analysis use three concrete examples to further restrict the characterization of those programs that are adapted for RCT geared at generating evidence for direct accountability purpose. We suggest to qualify these programs as “tunnel-programs” in that they satisfy the following requirements: (i) a period of observation coherent with the logical chain, (ii) a limited number of well-specified homogeneous inputs, (iii) a tried and tested process, (iv) a short and external events-proof causal chain, (v) a rapid and stable appropriation by beneficiaries, (vi) a large and stable participation, and (vii) a set of measurable outcomes in the short run, covering the mains aspects of the program.

Accountability purpose is an important aspect of evaluation and there is great expectation that impact evaluation could be more decisive in this field, particularly amongst donors. In this way AFD’s experience could be considered as a limitation of accountability potential of IEs and as an invitation to reduce this expectation.

Nevertheless, our purpose is not to dismiss the fact that accountability is only a part of the multiple objectives pursued with the development of impact evaluation, for example to provide robust and cumulative evidence on development policies (on pricing, access, etc.), to help foster the missing links between academic research and development practices, to promote trial and errors process in policy building with experimentation preceding scaling-up, to be more influential on policy makers, to test theories and to learn about households behaviours. In fact, Al Amana and Sky evaluations that are discussed in this article for their limitations in providing summative assessments of the program’s impact, nevertheless provide interesting results such as the first rigorous analysis on adverse selection in poor countries or a precise mapping of the very contrasted credit take-up rates in different areas of rural Morocco.

And in fact, impact evaluation specialists have generally turned towards the use of such methods towards field experiments geared at informing the design of future policies, rather than evaluations focused on existing ones. (Duflo 2009). But on the donor’s side however, accountability use of IE results remains an important motivation. It is this latter point that this paper has attempted to qualify.
References


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


