



Impact Evaluations of the Health Insurance Fund (HIF) projects

Jacques van der Gaag
Amsterdam Institute for International Development

July 2007

Wendy Janssens, Emily Gustafsson-Wright, and Tobias Rinke de Wit provided useful comments on an earlier draft, which resulted in many improvements of the exposition. Lucré Visser is responsible for the final editing.

Impact Evaluation of the Health Insurance Fund (HIF) Projects

Table of contents

1. Introduction
2. Why impact evaluation?
3. What is so special about impact evaluation?
4. Alternative set-up for impact evaluation.
5. The general set-up of impact evaluation studies for the HIF projects
6. What questions will be answered?
7. What questions can be better (and cheaper) answered in another way?
8. Conclusion

Annex

The Questionnaires used for the evaluation of the Okambilimbili project in Namibia.

1. Introduction

Global poverty is pervasive. Despite the fact that the world has more resources available than during any time in the history of mankind, more than one billion people eek out a meager subsistence on one dollar a day. And despite the efforts of the entire development community, which adopted the Millennium Development Goals (MDG's), the previous statement will, given current projections, be just as true in 2015, when the MDG's are supposed to have been reached, as it is today: we are continuously faced with the challenge to improve the living conditions of more than one billion people who suffer from extreme poverty and its correlates.¹

Poverty is more than a lack of money, or perhaps I should say it is less: poor people (as measured by income) also suffer from excess disease, illiteracy, malnutrition, lack of access to clean drinking water and sanitation, and lack of access to infrastructure such as roads to markets, electricity, legal systems and credit. The one solution that, over time, will reduce the number of poor and improve the living conditions of those at the bottom of the income distribution is sustained economic growth, as is clearly and unequivocally demonstrated by the success of the two emerging economic giants, China and India. Together they are home to more than a third of the world population. But their current economic success should not close our eyes to the hundreds of millions of people who still live in poverty there, and for the hundreds of millions more who live in countries that do not grow, or are even contracting.

In the absence of easy measures to reduce income poverty, many actors in the world of development have designed policies and programs to alleviate one or more of the correlates of poverty. Food subsidies, supplemental feeding programs, immunization programs, free elementary schooling, vouchers for secondary education, health cards, water projects, employment projects, micro credit, and rural development projects, are just some of the hundreds, maybe thousands, of different interventions that all aim at improving the living conditions of the poor or, even better, providing the poor with the tools to escape poverty.

Do they work? Which of the interventions are more successful than others? When new types of interventions are being introduced, will they be more or less successful than conventional projects? It is increasingly common in the development world to ask these questions, to ask them precisely, and to ask them before a program or project is being implemented. Indeed in many cases "impact evaluation", has become an integral part of project implementation, especially when the interventions are of an innovative nature.

¹ For an excellent book on this see *The Billion at the Bottom*, by Paul Collier, World Bank 2007.

One example of such an innovative project is the Health Insurance Fund (HIF) which aims at the introduction of low-cost health insurance for low-income workers in developing countries. In addition to providing the insurance, with a significant subsidy to make the premium affordable, HIF guarantees access to high quality and easy to reach medical care. HIF-sponsored projects provide a new way to increase access to good quality care for poor people. Because of its innovative character, and in line with current best practice in the developing world, it has been decided that all HIF-sponsored projects will be evaluated, using state-of-the-art impact evaluation techniques². This note will provide a brief general introduction to the why and how of impact evaluation. It will then sketch the specific design of the impact evaluation for HIF-sponsored projects and the questions this design will be able to address. Subsequently we will point to some issues that are relevant for measuring the success of the projects, but can be more easily addressed with alternative evaluation designs. The last section concludes.

2. Why impact evaluation?

Rather than giving a theoretical answer to the question “why is impact evaluation important?”, we will use an example that, in a short period of time, has become famous in the development world: The evaluation of The Progres-Oportunidades anti-poverty program of Mexico³.

In 1997, Mexico started an anti-poverty program, called Progres-Oportunidades, which enrolled 300,000 families in 12 states, at the cost of about US\$ 58 million annually. The program had many innovative and sometimes controversial components. It dispensed cash directly to poor households, rather than the customary in-kind benefits, but under the condition that households needed to change specific patterns of behavior⁴. Recipients must invest in health, nutrition and education, with a special focus on children. The amount of money followed a life cycle, with more money being received by households with many children, the transfers being increased if children progressed to secondary education, and the transfers being phased out when children became adults and moved out of the house. Perhaps most controversial was the fact that existing anti-poverty programs (mostly food subsidies) were being replaced by this new program. Not surprisingly, the constituencies of these programs (poor households and politicians alike) expressed heavy opposition to this new approach to fight poverty.

In addition to this immediate opposition, there was an additional thread to the sustainability of this program: the so-called stop-go problem. The stop-go problem refers to the fact that in Mexico, but by no means in Mexico alone, many good poverty alleviation efforts of the past that were launched by previous administrations, were interrupted by new administrations who preferred to put their own “brand name” on new programs.

² The Amsterdam Institute for International Development (AIID) and the Center for Poverty Related Communicable Diseases (CPCD) are charged with the evaluation of HIF-sponsored projects.

³ In this section I draw heavily on “*Progress against Poverty*” by Santiago Levy, Brookings Institution Press, 2006.

⁴ Hence the name “conditional cash transfer” for this type of program.

How could the skeptics of the new approach be convinced of the value of the program? How could future administrations be deterred from interrupting the program and replacing it by new ones?

To prepare for answering these questions, it was decided, at the start of the program, to put in place a rigorous impact evaluation system that would provide the evidence for the success (and perhaps failure) of the various program components that was needed to make informed decisions about continued program funding⁵.

By all measures, the evaluation program was state-of-the-art. Quantitative and qualitative analyses were performed. Data were collected on participants and on control groups. The same households were followed for a number of years. Well-known international scholars applied modern econometric techniques, and the results were published in peer-reviewed journals. Data were made available to other groups of scholars who could independently verify the results.

When president Fox took over from president Zedillo, under whose leadership the Progresa program had been implemented, he called in the group of scholars who had worked on the evaluation of the program. They reported on a large number of results, including the effect of the cash transfer on consumption, savings and investments, use of health services by all household members, children's health and nutritional status, school enrollment and school progression, and adult labor force participation.

On the basis of these results, president Fox decided to continue the program⁶. By 2005 the program was reaching 5 million households in 31 states, at an annual cost of almost US\$ 3 billion.

It would go too far to claim that the continuation of the Solidaridad-Progresa program was solely due to the program evaluation. But it is probably true that without the evaluation, political arguments for and against the program would have won it from unsubstantiated claims of success or failure. The available hard evidence of the impact of the program has been enormously influential. Indeed, it is currently being used to support the implementation of many new conditional cash transfer programs all over the developing world.

⁵ the same evaluation program would, of course, provide valuable information during the implementation of the program. Over time, many improvements were made in the program based on these evaluation outcomes.

⁶ but under a new name: *Oportunidades*.

3. What is so special about impact evaluation?

Programs and projects have almost always been evaluated, so what is so special about impact evaluation? Up to now, the common approach to project evaluation is to see whether the project has achieved its objectives in terms of activities and output (e.g. the number of water pumps built). Impact evaluations go a step further and evaluate whether these activities can be shown to have led to the desired outcomes (e.g. improved health).

As it turns out, there are many pitfalls for impact evaluation, and the solutions to many of those difficulties are highly technical. An excellent reference is Ravallion (2001)⁷. More technical treatments can be found in Heckman et al. (1999)⁸ and Blundell and Dias (2000)⁹.

In the so-called “hard” sciences, such as chemistry or physics, experiments or measurements can be carried out exactly as the scientists designed them, in the laboratory (or in particle accelerators, or in the universe), without third party (human) intervention. The experiments are conducted with inanimate objects and can be repeated endlessly under the same circumstances. Once a result has been obtained (and is repeatedly found to hold) it becomes a “law”. For instance, nothing can move faster than light, no matter how you measure it, and under what circumstances.

In the social sciences we are not so lucky. Much of the research is context-specific. What works in one setting, say a country, could prove less successful if replicated in another country. Moreover, it is impossible to “control” all relevant variables in the real world, as is done in a laboratory. Consequently, we have to be careful when making policy recommendations based on the success of one intervention in one particular environment.

Perhaps more importantly, trying to answer the basic question of “does it work, here and now?” i.e. ignoring the contextual situation, is already wrought with pitfalls. Here is an example (relevant to the HIF projects):

Health economists have long tried to answer the following question: *what is the price elasticity of medical care?* It seems simple enough. One could, for instance, measure the medical consumption of a group of households who have extensive health insurance coverage (and thus face a low price, or even a zero price, when visiting a doctor or when buying drugs), and compare this with the medical consumption of less well-covered households (who face positive prices). The difference in consumption will give the information on the price effect¹⁰.

⁷ Ravallion, Martin (2001), “*The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation*”, World Bank Economic View, vol. 15 (1), pp 115-140.

⁸ Heckman, James J., Lalonde, Robert J. and Smith, Jeffrey A. (1999), *The Economics and Econometrics of Active Labor Market Programs*”, in: A. Ashenfelter and D. Card (eds), “*Handbook of Labor Economics*”, vol III, Amsterdam; Elsevier Science BV, pp 1865-2097.

⁹ Blundell, Richard and Monica Costa Dias (2000), “*Evaluation Methods for Non-Experimental Data*”, Fiscal Studies, vol 21 (4), pp 427-468.

¹⁰ We assume, as is usually the case, that we can control for all other relevant factors such as family size, age, income, education levels, distance to a clinic, etc.

Not so. It is very likely that, for instance, young households who feel (and often are) healthy, who think that their risk of falling ill is low, and who expect not to have to visit the doctor often, decide not to buy (comprehensive) health insurance. We will observe little health care consumption under less than comprehensive coverage, but the causality runs from the former to the latter.

Alternatively, hypochondriacs may decide to obtain the best health care coverage money can buy, and visit the doctor at least once a week, whether they need to or not. Again, the causality runs from (expected) medical consumption to extensive coverage, and not the other way around¹¹.

The main point here is that when human actions, or *choices* (such as the voluntary choice to buy health insurance or not), are involved, the analysis which in first instance looks so simple, becomes more complicated.

This is not an isolated case. In economics, many of the outcomes we are interested in are the result of more or less voluntary choices by all the actors relevant to the issue in which we are interested. For example, try to answer the question whether men earn more than women, for the same type of work. The only observations we have, by definition, are wage rates for both men and women *who have chosen to join the labor market and who have received a job offer and have accepted it*. That is hardly a random sample of all men and women who could choose to join the labor market, and, if they did, would get a job offer that they find attractive enough to accept. Furthermore, the choice to accept a job will be influenced by the variable of interest, the wage rate. If there is gender discrimination in the labor market, many qualified women may have decided (chosen) not to join the labor force, which may seriously bias any inference from the difference in wages between men and women who are in the labor market¹². To be precise, one can always say, on the basis of data from *working* men and *working* women, whether *working* men earn, on average, more or less than *working* women. But one cannot generalize to a larger sample: i.e. to all (potentially working) men and women.

For impact evaluations, this problem always needs to be addressed: we can easily show whether people with health insurance consume more (or less) medical care than people without health insurance. However, if health insurance is made available, some people may decide (choose) to (voluntarily) participate in the scheme, and other may prefer to pass up the opportunity. These two groups are no longer random samples of the general population. They have *selected* themselves to belong to one group and not to the other. This leads to “selection bias”, which greatly complicates impact evaluations of programs or interventions in which individuals or households are free to participate or not.

¹¹ We will discuss a solution to this problem in the next section.

¹² This example was first addressed by Jim Heckman, who designed an econometric solution for this so-called selection bias. The solution is called the Heckman correction. He received the Nobel prize for Economics for his work on impact evaluation.

Let's look at another example: we want to find out whether professional training will help to get high school drop outs a better job. Unless you can force high school dropouts to take that training, you run into the same problem: you may find that those who voluntarily chose to take the training may be more successful in obtaining a job than those who chose not to enroll in training. But is that due to the specific training, or were those who decided to enroll in the training already more motivated to find a job than those who did not enroll? The problem here is that we can observe enrollment (which may not be the causal factor in obtaining a job), but we cannot observe "motivation" (which may be the causal factor).

In the medical world, the effectiveness of drugs is being tested in double blind trials. Patients receive a drug or a placebo from a doctor, and neither the doctor nor the patient knows who gets what. The patients are randomly selected in one or the other group. Let's assume that all patients have first been carefully chosen on the basis of precise diagnoses, and are subsequently randomly selected in the treatment or the control (placebo) group. Let's also assume that taking the drug at the right time, under pre-specified conditions ("before dinner, with a glass of water") is strictly enforced. Such an experiment would allow you to determine whether the chemical content of the drug has the desired effect.

But even such an experiment is impossible, or perhaps even undesirable, in the social sciences. If the same drug would be put on the market, what can the double blind experiment tell us about its effect *on the population*? Many things can go wrong: patients who, based on medical diagnoses (if they come and see a doctor) should get the drug, may be overlooked. Adherence to the drug regime may be less than perfect. Patients may change their behavior in other (as of yet unknown) ways that may enhance or reduce effectiveness. In other words, when behavior intervenes, impact evaluation becomes complicated. The effectiveness of a drug in the real world can be very different from the effect measured under laboratory conditions.

Impact evaluations are very different from descriptive analyses that report on the progress of a program. Such studies show how many people are enrolled in a program (e.g. health insurance), how often they visit a doctor, how much drugs they use, how less often they go on sick leave, etc. Impact evaluations will report on those aspects (and others, see section 6), but always *relative to what would have happened in the absence of the program*. The world without the program (the placebo) is the proper counterfactual against which we need to measure the success of a program. It is unfortunate that in many cases the real counterfactual is impossible to observe. The challenge is to construct a control group that resembles, as well as possible, the missing counterfactual.

These are only a few examples of some of the things that can go wrong with impact evaluations. In the next section we will discuss the various problems that need to be overcome before we can state the effect of program impact with sufficient confidence. We will also list, and briefly describe, the various techniques that exist to overcome these methodological problems. These techniques call for various evaluation designs. But we will also see that every evaluation design comes with its own set of problems. In

subsequent sections we will describe the proposed design for impact of evaluation of the projects sponsored by the HIF, and the questions we will be able to answer with it.

4. Alternative set-ups for impact evaluation

Let's try once more to measure the effect of different prices on the use of medical care. As we saw above, it is invalid to gauge this price effect from a comparison of medical consumption by households who face high prices (because they are not insured) with households who face low prices (because they are insured). And the reason is that people *chose* their health insurance coverage freely, in anticipation of future health care needs. In economic jargon: health insurance coverage is *endogenous*.

What can be done about this? In an ideal world (from the point of view of a researcher), households would be given various types of health insurance, with various levels of coverage, and with various co-payment and deductible features. The health insurance packages would be distributed at random, and the households would not be able to refuse them or alter them. In other words, health insurance coverage would no longer be a matter of choice. It would be *exogenous*. If we now find differences in health care utilization among the various levels of insurance coverage, we can be sure that these differences are *caused* by the deductibles, co-payments and coverage levels in which these packages differ. It seems farfetched, however, to force households to accept a certain insurance package ("whether they like it or not") just to be able to estimate the price elasticity of medical care. Still, this has been done.

In 1971, the RAND Corporation started a large study, the Health Insurance Experiment (HIE), aimed at answering the following three questions:

-How does cost sharing or membership in an HMO affect use of health services compared to free care?

-How does cost sharing or membership in an HMO affect appropriateness and quality of care received?

-What are the consequences for health¹³?

The HEI was a large-scale randomized experiment conducted between 1971 and 1982. For the study, RAND recruited 2,750 families encompassing more than 7,700 individuals. Participants were randomly assigned to one of five types of health insurance, with various levels of cost sharing, from zero percent (no co-payments) to 95 percent. Out-of-pocket spending was capped so as not to exceed various percentages of income.

A brief summary of the findings include:

¹³ For more detail on the RAND Health Insurance Experiment, see the RAND HEALTH website (www.Rand.org/health).

- Participants who paid for a share of their health care used fewer services than a comparison group given free care.
- Cost sharing reduced the use of both highly effective and less effective services, but did not affect the quality of care.
- Cost sharing in general had no adverse effect on participants' health¹⁴.

Randomized social experiments have been conducted in the US to evaluate a wide variety of public policies and programs, from welfare reform to the Negative Income Tax. Randomization is increasingly used in developing countries to evaluate anti-poverty programs (such as vouchers for secondary education, or supplemental food programs). But randomization is costly, because households need to be enticed to participate (why else would they accept an insurance policy that they would not choose voluntarily?)¹⁵. When conducted in the area of health care, it is also riddled with ethical issues. What do you do when a household that has randomly been assigned to an insurance package with high deductibles becomes too poor to buy necessary health care? But from an evaluation point of view, they are considered to be the golden standard against which all other methods need to be compared.

Which other, non-experimental methods are there? The following table is adopted from a report titled *The Impact of Health Insurance on Access, Utilization and Health Status; The Case of Colombia*¹⁶.

It shows three alternatives to randomization. The first one, "Matching" is an econometric technique in which participants (say, in health insurance) are statistically matched with non-participants on the basis of measured characteristics (income, age, sex, education, etc.). By applying this method it is assumed that all determinants of participation can be satisfactorily captured by the characteristics available in the data. The possible impact of unmeasured characteristics (such as motivation or hypochondria) can not be accounted for.

The next alternative is called the "Double Difference" method. In essence, for this method two groups of households (one with and one without insurance) are followed over time. The *differences* between these two groups in the *changes* over time in health care consumption form the basis¹⁷ of the impact evaluation study.

The third method in the table, "Matched Difference", is a combination of the previous two and combines the advantages of both.

¹⁴ In the research reports of this study, these results are, of course, all quantified.

¹⁵ Households also need to be enticed to continue to participate, and to fill out many questionnaires on a regular basis. Indeed it is not unusual that people drop out of such social experiments, which leads to so-called attrition bias, the flip-side of selection bias.

¹⁶ Giedion, Ursula, Beatriz Yadira Diaz and Eduardo Andres Alfonso, *The Impact of Health Insurance on Access, Utilization and Health Status: The Case of Colombia*, draft report, mimeo, November 2006.

¹⁷ See next section.

The final method is one of a number of econometric techniques that try to deal with the endogeneity problem.

All the methods have advantages and disadvantages. In practice, a number of methods will be used simultaneously, to test whether the results obtained are sensitive to the methods used. To make this possible, the evaluation set-up needs to be sufficiently broad and flexible. We will discuss the organization of the evaluation for the HIF-sponsored projects in the sections below.

Some alternative methods for project evaluation

Nº	Method	Advantages	Disadvantages	Required data
1	Randomized trials	Technically the first best option and constitutes the benchmark for all other quasi experimental methods.	May present problems of internal and external validity.	Data from experimental design.
2	Matching	Eliminates the selection bias related to observable characteristics.	Does not control for unobservable selection bias. A problem of a limited area of common support may exist.	Cross section data.
3	Diff-Diff	Eliminates selection bias due to unobservable characteristics.	Assumes time invariant selection bias. Does not control for time variant unobservables.	Repeated cross section or panel data.
4	Matched double difference	Controls for observed and unobserved selection bias.	Assumes time invariant selection bias. Does not control for time variant non observable characteristics.	Repeated cross section or panel data.
5	IV	Corrects for endogeneity problems.	It is often hard to find an instrumental variable that substantially affects participation and is convincingly unrelated to outcome. Calculates the marginal effect of treatment.	Cross section.

Source: adopted from Giedion et al. 2006.

5. The general set-up of impact evaluation studies for the HIF projects

When designing an impact evaluation program for HIF-supported projects, we have to be practical. The so-called gold standard for impact evaluation, i.e. random assignment of the intervention, would greatly complicate project implementation. Some people will get

health insurance, and access to health care, but their neighbors will not¹⁸. We do not opt for this design.

The second best, and most often used design makes use of treatment and control groups, follows these groups over time, say 3 or 5 years, and analyses the impact of the program on the basis of so called double-difference method¹⁹. This is pretty straightforward:

In a first stage, the null-measurement, data are collected on both the treatment group (i.e. the group eligible for the intervention) and a control group. Let's call the treatment group A, and the control group B. The measurements from the baseline survey may result in A0 and B0. A0 and B0 could be the total out-of-pocket health expenditures in the year prior to the baseline survey, for group A and B, respectively. If the control group is properly chosen, A0 and B0 will be the same on average.

After one year the measurements are repeated, but group A has had the benefit of being eligible for health insurance. Group B has not. Let's denote the new measurements A1 and B1.

Clearly A1 will differ from A0 because the A group is covered by health insurance. However, it is quite possible that B1 also differs from B0. After all, a year has passed, and a lot can happen in a year. The government may have started a health promotion campaign. The general economy may be growing rapidly, thus allowing households (even those who are not insured) to spend more on health care. Or the country may have experienced a drought or other natural disaster, which has reduced the income levels of everyone (and thus reduced health care expenditures across the board). Whatever the case, changes in health expenditures are not influenced by health insurance alone, and we have to control for the effect of those other factors.

The double difference method deals with these intervening factors as follows:

Let $DA1 = A1 - A0$ be the difference in health care expenditures of group A, between year 1 (with insurance coverage) and year 0 (no coverage).

Let $DB1 = B1 - B0$ be the difference in health care expenditures of group B, between year 1 and year 0 (both years no coverage).

Then the double difference $DD1 = DA1 - DB1$ is the correct measure of the impact of the introduction of health insurance. It is the *difference* in the *change over time* of health care expenditures by group A and group B, due to the introduction of health insurance in group A.

¹⁸ Apart from logistical problems, there are many ethical issues here.

¹⁹ Also called the differences-of-differences method.

Since both groups, by design, have equal chances of being affected by the same external factors (income growths, droughts, new government policies, etc), the remaining difference can be attributed to the one factor in which they differ: eligibility to health insurance.

In sum, for the impact evaluation of HIF-sponsored projects, we will always take the following steps:

- define a treatment group and a control group;
- do a baseline survey on both groups (the null measurement);
- repeat the survey on the same households for a number of years (to create a set of panel data);
- use the double difference method as the basis for the analysis.

The double difference method becomes more powerful over time when the long-term impact of the intervention becomes apparent, because, given the longer time span, many other variables may have intervened. In addition, as mentioned above, the method can be combined with other methods (such as matching techniques) to provide additional tests for the validity of the evaluation results.

6. What questions will be answered?

Up to now we have discussed a number of issues that need to be addressed for impact evaluation, and that have direct consequences for the general structure of the evaluation study. We have not yet discussed what type of data we will collect on both the treatment and the control group. The key words in this section will be “comprehensiveness” and “flexibility”.

During the seventies and early eighties, the United Nations were very active in setting up programs in developing countries to systematically collect household data on a large variety of topics. Under the guidance of the United Nations Household Survey Capability Program, developing countries implemented income and consumption surveys in one year, education surveys in the next, and employment, agricultural and a variety of other surveys in subsequent years. Once these surveys had been completed, the cycle of surveys was started again. These surveys were excellent in measuring precisely the level of school enrollment, or health care consumption, or income inequality, or cassava production. They were less useful, however, if one wanted to answer such questions as what causes low school enrollment, or poor access to health care, or high income inequality or low cassava production.

In order to answer such questions, the various types of information needed to be collected together, at the same time, and for the same households. For instance, when analyzing the causes of low school enrollment one needed to know the income level of the parents, the number of children in the household, the education level of the parents, whether or not children participated in the labor force to contribute to household income, the children’s

health and nutritional status, the distance to the school, the quality of the education received at that school, and a host of other variables that are relevant to the households' decision to invest in the human capital of their children.

These types of considerations were behind the World Bank's decision to invest in the development of a comprehensive household survey that included a large number of variables that were all thought to be relevant to the welfare for households in developing countries. These multi-purpose surveys, called Living Standard Measurement Study (LSMS) surveys, have been implemented in dozens of developing countries since 1985²⁰.

The questionnaires for the evaluation of the HIF-sponsored projects are adaptations of these multi-purpose LSMS surveys. The survey always includes a set of standard modules (described below) and a set of modules that are specially designed for the purpose of the study. This flexibility in the design of the questionnaire makes it an ideal instrument for a large variety of impact evaluation studies.

In what follows, we will briefly describe the contents of the questionnaire used for the evaluation of the Okambilimbili project in Namibia. As in the HIF-sponsored projects, the Okambilimbili project aims at providing low-cost health insurance for low-cost workers. Experience with the evaluation of this project is thus extremely valuable for the HIF evaluation efforts.

The following is a list of standard modules that are always included in the questionnaire²¹:

Household roster

this module determines the composition of the household by collecting, on each member, data on age, sex, and relation to the head of the household.

Employment and income from work

employment and unemployment data are collected, including information on type of work, income and other benefits.

Housing conditions

rental versus ownership, size of dwelling, type of water supply, sanitation, roof condition, electricity.

Consumption expenditures

this module needs a bit more explanation. The above modules are all closely related to the measurement of household welfare (or poverty). Consumption

²⁰ Warning, the author was involved in the development of the first LSMS survey, in 1985, in Cote d'Ivoire, and might thus be biased. For a comprehensive description of LSMS surveys, see "Designing Household Survey Questionnaires for Developing Countries; Lessons from 15 years of the Living Standards Measurement Study", Margaret Grosh et al., The World Bank; 2000; three volumes.

²¹ This is not a complete listing. A copy of the actual questionnaires used for the baseline survey in Namibia is attached.

patterns are of interest by themselves, but in this type of survey the main reason why we collect consumption data is that we need a reliable monetary measure of welfare. Income would be the ideal measure, but in situations where regular income (such as monthly wages) is a rarity, it has been argued that consumption (suitably adjusted for family composition) is the preferred monetary welfare measure. Nevertheless, the consumption data do allow for the analysis of food consumption, non-food consumption and durables, in some detail. Of course, for the study of the impact of health care insurance, data on the consumption of medical care and services need to be collected in great detail.

Since the focus is on health and health care, the following modules are expanded versions of the standard ones:

Access to various forms of health insurance

in addition questions are asked on the willingness-to-pay for health insurance (if it would become available).

Health status

the latest version of the LSMS survey includes a large section of self-reported health status. That expanded health module is included in the Namibian questionnaire. But for the current purpose, that was not enough (see below).

Health expenditures

detailed data on health care expenditures (clinic visits, drugs, hospital visits, and associated costs, such as travel costs) are being collected; in addition information on health insurance coverage and out-of-pocket expenses is included.

Finally, after visiting the households for the first time (with a number of follow-up visits to complete the questionnaire), the households are visited for a second time, to collect even more detailed health data. These data include:

Health status

More detailed data on self-reported health status are being collected.

KAP data

KAP stands for Knowledge, Attitudes, and Practice. In this case it refers to HIV/AIDS risk factors.

Specific health measurements

Because of the special focus of the Okambilimbili project on coverage for HIV/AIDS support and treatment, the second part of the survey is carried out by specially trained nurses. After asking a few simple questions about self-reported health status, the nurses will measure the blood pressure of each member of the household. Subsequently, the nurses ask permission to apply a non-intrusive

saliva test which will determine the HIV status of the household members. Written consent is required for this test.

In sum, for the evaluation of the impact of the Okambilimbili project on the well-being, broadly defined, of participation households, we use a comprehensive and flexible approach to data collection: the LSMS survey, adapted for the specific purpose at hand.

The questions we will be able to address are wide-ranging. They can be organized in three groups:

Process questions, such as: who enrolls in the insurance and what are the factors determining that decision?

First impact questions, such as: does enrollment increase health care utilization, by how much, by which participants (adults, children, by education level, employment level, etc.), what type of health care? Does it reduce out-of-pocket expenditures? Is it successful in protecting participants from falling below the poverty line?

Longer term effects, such as: What happens to the health status of participants, relative to non-participants? But we can also look at changes in behavioral patterns as a result of the elimination of the risk of having to deal with large and unexpected health expenditures. Some of these effects can, more or less, be predicted, but still need to be quantified. For instance, research suggests that a reduction in the volatility of welfare (income, consumption) will allow households to invest more in the human capital of their children. Other effects

For instance, in the PROGRESA program, it was found that recipients did not use the entire cash transfer to increase consumption; some of it was saved. After a couple of years, the savings were large enough to buy productive assets, such as a cow, or a tractor, or other tools. When the cash transfer stopped (because the children had left the house), the new productive assets allowed the households to earn a higher income than before the start of the program. Because it is virtually impossible to predict such effects in advance, the data collection needs to be both comprehensive and flexible.

7. Which questions can be better (cheaper) answered in another way?

Despite the comprehensive data collection effort, there are many questions that will come up during the course of project implementation. Some of these questions can be added to the household questionnaires²², other questions need to be addressed in a different way.

²² About one year after the baseline survey, the same households will be asked to fill out the same questionnaires. It is very easy, and virtually without costs, to add a few questions to this questionnaire that may have come up during the first year of implementation.

For instance, if we are interested in how satisfied participants are with the insurance program, or with the services in a particular clinic, it does not make any sense to ask the entire sample to answer this question. After all, only a small percentage of the population sample will actually participate in the program, or visit that particular clinic. It would be much better, and cheaper, to get a list of all program participants, and ask them these (and similar) questions, using a brief, perhaps one page, questionnaire.

But we may become interested in the population's satisfaction with the quality and accessibility of health care in general. That would be a good topic to address in the general survey. Doing so would also allow a comparison between the levels of satisfaction reported by the project participants and by those who are not (yet) enrolled.

Another set of issues need to be addressed at the level of a clinic. We could ask (and often do), in the general survey, whether the doctor is always available, or whether the necessary drugs are always present. But if we want to know this at the level of participating clinics, it is much better (and cheaper) just to go to those clinics and collect the information there.

8. Conclusion

In the social sciences, we will never be able to create the type of laboratory conditions that physicists or chemists enjoy. It is quite common that, during the implementation of a project, changes are made in various components of the intervention, or in the groups that are eligible to benefit from the intervention, or in other program characteristics. Indeed, such changes are often made as a result of program evaluations²³.

Trying to evaluate programs or projects that keep changing over time is like trying to shoot a moving target. It would be highly undesirable, however, if the need for improvements of the program would be overlooked, just to facilitate a proper evaluation. The needs of the participants will always trump those of the evaluators.

In a few cases the evaluation can actually suffer from the success of a program. The evaluation of the Okambilimbili project in Namibia, mentioned above, is such a case. The AIID has received a grant from the Dutch Government to evaluate the impact of this project. Given the relatively small size of the project, it was decided to define a treatment group (those who were eligible to participate in the insurance) and a control group (those who could not know about it, for instance because they lived in another part of the city). The project managers developed the first low-cost insurance package in co-operation with just one insurance company. When other companies learned about this, they too wanted a part of the low-cost end of the market. As a result, they developed their own low-cost insurance products. Before the intervention, not a single low-cost insurance package was available for low-income workers. At the time of this writing a number of different low-cost insurance products are on the market. By this measure alone, the

²³ though they are often due to political interference, or just a change in manpower for the management of the program.

Okambilimbili project is a huge success. But this success strongly interfered with the evaluation: it effectively wiped out the control group consisting of people who could not know about the availability of low-income health insurance. Everybody now knows about it and can choose to enroll.

Fortunately, there are numerous ways to adjust the evaluation design²⁴. But the example should serve as a warning that, despite the best of our intentions, the science of impact evaluation of interventions that are being implemented by and for people, is an art.

²⁴ For instance, to focus on before/after evaluations, or to over-sample households who become insured because their company has decided to participate in the new low-cost insurance programs.

