INTER-AMERICAN DEVELOPMENT BANK
Social Development Department
Poverty and Inequality Advisory Unit

Technical Note 2
December 1, 1999

Poverty & Inequality Technical Notes

IMPACT EVALUATION METHODS FOR SOCIAL PROGRAMS

Ferdinando Regalia[*]

This technical note provides a brief overview of the most commonly used methods to carry out impact evaluation of social programs.

Why Evaluate a Project's Impact?

Impact evaluation is an indispensable tool to assess whether a program is achieving its objective, how the beneficiaries' situation changed as a result of the program and what the situation would have been without the program. Moreover, if an impact evaluation is carried out at an intermediate stage of project execution, very important lessons can be learned on how the program design and/or the project execution can be modified to improve the effectiveness of the intervention. The definition of program objectives and targeting mechanisms might also be improved by planning an impact evaluation at an early stage of project design. While the design of an impact evaluation can be time and resource intensive, the costs are very often small relative to the scale of a transfer program (particularly if in-country resources in terms of available data and data processing skills are used). The returns in terms of increased effectiveness of social spending and greater accountability are very high.

Evaluation Methods

Impact evaluation tries to answer the question: What would have happened if the program had not existed? All impact evaluation methods compare a treatment group (the program beneficiaries), with a control group of non-beneficiaries. These methods fall into two categories: *experimental* and *quasi-experimental designs*. These methods assume that the programs will not impact general conditions in the economy. Indeed, general equilibrium effects of small-scale programs should be insignificant. Evaluation of large-scale programs should account for general equilibrium effects.

A Experimental Design

Randomization. This is the most robust of all the evaluation methodologies. Once the target population is chosen on the basis of observed characteristics, the program's actual beneficiaries and non-beneficiaries are selected randomly within the pool of eligible beneficiaries. By definition randomization implies assigning eligible beneficiaries to a treatment and a control group through a lottery. Randomization ensures that there are no systematic differences in the observed characteristics between program participants and individuals in the control group. The impact of the intervention is assessed by subtracting the mean outcomes of the group of beneficiaries from the mean outcomes of the non-beneficiaries in the control group. This can be done for any indicator of interest (income, consumption, school attendance, labor force participation, etc.). Randomizing beneficiaries is feasible (and ethical) whenever budget constraint require rationing of program benefits. Even when the programs are national in scale and aim at 100 percent coverage, expansion of coverage is often gradual, and randomization offers an ethically sound basis on which to proceed (since all targeted individuals have the same probability of being selected). Individuals who function as controls at an early stage of program implementation become beneficiaries at

[*] Ferdinando Regalia is a Consultant in the Poverty and Inequality Advisory Unit, IDB.

a later stage. An example of an experimental design impact evaluation system is PROGRESA in Mexico, a targeted human development program.

B Quasi-experimental Design

Quasi-experimental methods construct control groups that resemble treatment groups through econometric techniques and not randomly by means of a lottery among eligible beneficiaries. Quasi-experimental methods can make use of existing data. They require the existence of a survey administered to both beneficiaries and non-beneficiaries of a program. Although they may not ensure the same level of reliability of results because they cannot fully control for selection bias, these methods are in general less costly to implement. Using a combination of quasi-experimental methods helps control for selection bias.

Matching. This method builds a control group by selecting among the respondents of a large-scale, often national, survey those individuals whose observable characteristics are similar to those of the program beneficiaries. When the comparison includes a large set of observed characteristics, the matching between beneficiaries and non-beneficiaries can be performed using propensity scores. Propensity scores measure an individual's predicted probability of being a program participant given her observed characteristics. Propensity scores are usually obtained from the estimation of binary choice nonlinear econometric models using the whole sample of beneficiaries and non-beneficiaries. The matching method pairs participants and control group members from a similar socioeconomic environment with the closest propensity scores. A measure of this closeness is the absolute difference in scores. The impact of the intervention is evaluated by subtracting the mean outcomes of the group of beneficiaries from the mean outcomes of the matched non-beneficiaries belonging to the control group. The *Matching* method is useful to carry out an impact evaluation when no baseline data have been collected before the program implementation. The evaluation of the Argentinean *Trabajar*, a workfare program, was carried out using a propensity score method. The results of the program evaluation depend on the set of observable characteristics used to compute the propensity scores. Significant differences in the distribution of observable characteristics between the control and treatment groups might bias the results. Accurate weighting of the two groups helps to reduce this bias. A second source of bias is more relevant. It arises when unobservable individual characteristics systematically influence both the program participation and the outcome variables that are the object of the impact analysis, i.e. selection bias. Programs that use self-selection targeting criteria, such as workfare programs, might be particularly subject to this second sort of bias.

Reflexive comparison. This method requires a baseline survey of program beneficiaries before the program is implemented and a follow-up survey. The baseline represents the control group, and the evaluation is performed by comparing the average change in outcome indicators before and after the intervention. This method however cannot identify the impact of the program from that of other factors (e.g. economy wide changes) that have affected the beneficiaries. For this reason results are biased, and the direction of that bias is difficult to assess.

Difference in difference. This method can be used to reduce the potential selection bias (when unobservable individual characteristics are assumed to be time invariant) and the impact of other factors exogenous to the program on observable characteristics. It accomplishes this by looking at the difference in outcome of participants relative to the difference in outcome of non participants. Equivalently, it looks at the difference in indicators for the two groups at the end of the program relative to the difference in indicators at the beginning. Let X be the indicator of interest, and the subscript T and C indicate treatment and control groups, and time index 0 and 1 indicate the time before and after the implementation of the program, then this method computes the following double difference:

$$D = (X_{T1} - X_{C1}) - (X_{T0} - X_{C0})$$

Regression analysis allows for controlling of differences in initial observed characteristics between control and treatment groups and for changes in exogenous variables.

Regression methods based on instrumental variables. Sometimes it is neither possible nor desirable to do a baseline and follow-up survey, particularly when households originally included in the baseline survey are likely to drop out from the sample non-randomly (attrition bias). When outcomes are observed both for participants and non-participants after program implementation, instrumental variables can be used to evaluate the program impact without incurring problems of selection bias. Any variable that is correlated with individual participation in the program, but is non-correlated with individual outcomes given participation, can be used as an instrumental variable. This method is carried out in two steps. First, participation in the program is predicted using instrumental variables. Then, mean outcome indicators are compared conditional on predicted participation and nonparticipation. Finding appropriate instrumental variables is often very hard, making the implementation of this method rather difficult.

C General Equilibrium Effects

All the above evaluation methods assume that the programs have no effect on non-participants. In other words, these methods rest on two very strong assumptions that are not always satisfied. First, that the distribution of individual outcomes within the control group of a given program can be used to approximate the distribution of individual outcomes if the program did not exist. Second, that the distribution of individual outcomes within the treatment group of a given program can be used to approximate the distribution of individual outcomes if the program is universally applied. The first assumption is plausible only if the program size is small and all the general equilibrium effects generated by the program, inclusive of taxes and spillover effects on factor and output markets, are considered to be insignificant. The second assumption implies that it is reasonable to forecast the outcomes of a program's expansion by relying on the results of an evaluation carried out on a very reduced size program. However, this is not always the case because the expansion might give rise to important general equilibrium effects. Such general equilibrium effects should therefore be taken into consideration when fully assessing the impact of a program and when carrying out a rigorous cost-benefit analysis. This, however, is not an easy task.

Further Readings

Baker, J. L. (1999) "*Evaluating Project Impact for Poverty Reduction: A Handbook for Practitioners", *Mimeo. LCSPR/PRMPO. World Bank Washington D.C.

Gómez de León, J. and S. Parker (1999), *"The impact of anti-poverty programs on labor force participation: the case of Progresa in Mexico"*, Mimeo. Progresa team, Mexico D.F.

Heckman, J.J., Ichimura, H. and P. Todd (1997), "*Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program"*, Review of Economic Studies, October.

Heckman, J. J. and J. Smith (1998) "*Evaluating the Welfare State"*, Frisch Centenary Econometric Monograph Series, Cambridge, Cambridge University Press.

Jalan, J. and M. Ravallion, (1999) "*Income Gains from Workfare and their Distribution."*, Mimeo. Washington D.C.

Ravallion, M. (1999) "*The Mystery of the Vanishing Benefits: Ms Speedy Analyst's Introduction to Evaluation"*, Mimeo. World Bank Washington D.C.

Schultz, T. P. (1999) "*Preliminary Evidence of the Impact of PROGRESA on school enrollments from 1997 and 1998"*, IFPRI Report. International Food Policy Research Institute, Washington D.C.

Resources

External
Paul Gertler, U.C. Berkeley
James J. Heckman, University of Chicago
Martin Ravallion, World Bank
Petra Todd, University of Pennsylvania
Abt Associates Inc., Bethesda, MD, http://www.abtassoc.com

IFPRI, International Food Policy Research Institute, Washington D.C., www.ifpri.org

IDB
Carola Alvarez
Omar Arias
Arianna Legovini
Gustavo Marquez
Carmen Pagés-Serra
Ferdinando Regalia