World Bank
Employment Policy Primer
December 2002 ■ No. 2

Impact Evaluation

*Techniques for Evaluating Active Labor Market Programs**

Background

Over the past 40 years, "active" labor market programs (ALMPs) have emerged as an important employment policy tool. Their objective is primarily economic – to increase the probability that the unemployed will find jobs or that the underemployed will increase their productivity and earnings. ALMPs include job search assistance, training and retraining, and job creation programs (public works, micro-enterprise development, and wage subsidies). With economic reform, increasing liberalization of markets and growing concerns about the problems of unemployment, ALMPs have increasingly become an attractive option for policymakers.

Expenditure on these programs has, however, not increased substantially over the 1990s, remaining fairly constant at around 0.7% of GDP. This reflects to some extent the ambivalence of policymakers about the effectiveness of ALMPs. A frequently asked question is, "Are these programs effective?" Attempts have been made in OECD countries to answer this question through rigorous evaluations that compare outcomes for individuals who participate in the program (treatment group) with those of a similar group of individuals who did not receive the program (control group). However, such analysis has been mostly lacking in developing countries. Part of the problem lies in the lack of an evaluation culture in many countries, often due to low capacity for evaluation. Policymakers may not be conversant with the importance of conducting evaluations and the techniques used to conduct such evaluations.

There are many different types of evaluations:

- *process evaluations* focus on how a program operates and on activities undertaken in delivery;
- *performance monitoring* provides information on the extent to which specific program objectives are achieved (e.g. number of unemployed trained); and
- *impact evaluations* focus on the issue of causality to see whether a program has its intended impact (e.g. percent increase in employment and wages attributable to program) and which characteristics of the program led to the impact.

This note will focus on impact evaluations of ALMPs. It will discuss the objectives and importance of rigorous evaluations, highlight commonly used impact-evaluation techniques, and discuss who should conduct evaluations.

Uses of Impact Evaluations

The purpose of evaluations of ALMPs is to examine the effectiveness of programs against their stated objectives. On the basis of this, evaluation can then be used to:

- Help design new programs
- Refine program design
- Improve program targeting
- Identify ineffective programs

*This note was prepared by Amit Dar and edited by Tim Whitehead.

The World Bank Employment Policy Primer aims to provide a comprehensive, up-to-date resource on labor market policy issues. The series includes two products: short notes, such as this one, with concise summaries of best practice on various topics and longer papers with new research results or assessments of the literature and recent experience. Primer papers and notes are available on the labor markets website at <www.worldbank.org/labormarkets> or by contacting the Social Protection Advisory Service at (202) 458-5267 or by email at <socialprotection@worldbank.org>.

Box 1: Evaluation of Pilot Programs

Pilot demonstrations have proven extremely effective in the U.S. for testing new programs. Many policies and programs were first tested in a small number of sites before being promoted by policymakers for national implementation. Evidence from these pilot tests (usually based on experimental design evaluations) is often used to convince legislators to approve the national implementation of new programs.

One example of the successful use of experimental pilot demonstration in developing a new program is the U.S. Department of Labor's Self-Employment Demonstrations. An experiment was conducted at a number of pilot sites in two states. Unemployed individuals were assigned to a treatment group (that received new self-employment services) or to a control group (that did not receive such services). The quantitative evaluation results were so convincing that the U.S. Congress approved legislation to authorize the national implementation of a self-employment program.

Help design new programs. Ideally, policymakers should assess effectiveness of programs by first implementing and evaluating pilot projects. Evaluators might design a demonstration with one group participating in a program and a similar group of non-participants. Comparing the performance of the two groups over time would reveal the effectiveness of the program (Box 1). Based on these evaluations, policymakers can design and target programs more effectively.

Obviously, program efficiency can be (and is) regularly evaluated through the life of the program in order to:

Refine Program Design. In many countries, governments undertake rigorous evaluations of ALMPs to learn what works best so they can implement the most effective program design. For example, in Poland in the mid 1990s, public works were considered a costly intervention with few program participants going on to get regular wage employment. An impact evaluation found that a much higher rate of re-employment in non-subsidized jobs was achieved if public works were managed by private companies. This led authorities to change the design of the program – regulations were altered to favor private companies running public works projects – leading, over time, to improved cost-effectiveness of the program. While this particular effect is not generalizable across countries, it demonstrates the importance of conducting such evaluations.

Improve Program Targeting. Evaluations can enable policymakers to make informed decisions about which target groups benefit most from programs, resulting in targeted programs and enhanced program performance. For example, in the Czech Republic and in Turkey, an evaluation was designed to test the efficiency of vocational training for the unemployed. Evidence from the evaluation indicated that vocational training was more effective for women than for men – especially in relation to earnings. This led to the program being more tightly targeted towards women. It is worth noting that political and other considerations may dictate the ultimate decision on targeting of ALMPs. However, the responsibility of the evaluator is to carry out rigorous evaluations and present accurate findings to policymakers.

Identify Ineffective Programs. Some programs are ineffective and should be eliminated or changed; rigorous evaluations will help policymakers to identify them and allow resources to be redirected to programs that are more cost-effective. The evaluation of the Job Training Partnership Program Act (JTPA) program (Box 2) is one example of the use of quantitative evaluations to adjust budget allocations.

Impact Evaluation Techniques

Impact evaluations attempt to measure whether – and by how much – participants benefit from an ALMP. Outcome measures can vary depending on the evaluator's choice. The most common are earnings and employment rates, but evaluations have also been used to measure other employment-related outcomes.[1] While information on program participants is usually available, the challenge for a good evaluation is how to accurately represent the counterfactual – that is, how to construct an appropriate control group.

In many countries, the most commonly used evaluation technique is that which does not use a control group. These techniques rely instead on statistics compiled by program administrators (e.g. number of gradu-

[1] In many OECD countries, where these programs are offered as a substitute to (or even complement) welfare payments, outcomes are also measured in terms of savings in welfare payments and the likelihood of being on welfare. Some evaluations also attempt to measure social outcomes, e.g. changes in criminal behavior, drug use and teenage pregnancy.

Box 2: Eliminating Ineffective Programs

In 1986, USDOL initiated the National JTPA Study, a multi-year experimental evaluation of the effectiveness of programs funded by the Job Training Partnership Act. The study used a randomized experiment to estimate program impacts on earnings, employment and welfare receipt of individuals served by the program. This study produced one of the richest databases for the evaluation of training program impacts.

A rigorous evaluation of this experiment indicated that the program had very different results for adults and for youth. For adults, the program was successful in raising earnings by 7-11% and providing benefits of about $1.50 for every dollar invested. For youth, however, the program was not successful: there was no statistically significant impact on earnings, and costs exceeded benefits to society. These results clearly signalled that training services provided to youth were ineffective.

Following the release of the results in 1994, Congress cut the budget for the youth component of JTPA by more than $500 million (80%); the budget for the adult component was increased by 11%. By adjusting the JTPA budget among these components, Congress shifted funds from an ineffective program to an effective program.

ates, employment rate of graduates) or on the beneficiaries' assessment of programs. These evaluations are of little use. Without a control group, it is difficult to attribute success or failure of participants to the intervention, as these effects are contaminated by other factors, such as worker-specific attributes. Moreover, they don't control for how well the participants would have done in the absence of the intervention (Box 3).

In some cases, these evaluations provide information on deadweight loss, as well as substitution and displacement effects.[2] This may be useful in targeting programs towards certain areas/groups. Nonetheless, it is difficult to judge the robustness of the results, as this depends on how the sample was chosen and how respondents were interviewed. Hence it is more appropriate to conduct impact evaluations using techniques that use a control group.

[2] Annex 1 provides a glossary of commonly used terms in the impact evaluation literature.

Techniques that use a control group

These techniques are of two types: experimental and quasi-experimental. Experimental evaluations require selection of treatment and control groups prior to the intervention. In quasi-experimental studies, treatment and control groups are selected after the intervention. To compute program effectiveness, statistical techniques correct for differences in characteristics between the two groups.

Experimental Evaluation Techniques. This technique is based on the principle that, if large samples are randomly assigned to treatment and control groups, observable and unobservable characteristics of the two groups should not differ on average, and so any difference in outcomes can be attributed to program participation. The main appeal here lies in the simplicity of interpreting results – the program impact is the difference in the means of the variable of interest between the sample of program participants and control group. (For

Box 3: The Importance of Control Groups — A Hypothetical Example

In the town of Abca, 1,000 mineworkers were laid off as a result of the closure of the ABC Mining Company. Based on random selection, 500 were given a severance package while the other 500 were put through an intensive retraining program in computer skills. All 1,000 individuals were monitored over time.

Three months after the completion of the training, it was observed that 400 trainees were employed. This employment rate of 80 percent for the "treatment" group was touted by many as the impact of the training program.

However, Abcan evaluators cautioned against using only this figure to judge the success of the program. They wanted to compare this employment percentage to that of the "control" group – those who did not go through training. It was found that 375 of the control group of 500 were also employed three months after the "treatment" group completed its training – an employment rate of 75 percent. Hence, Abcan evaluators judged that the true impact of the training program was five percent and not 80 percent.

While this example makes many generalizations – no selection or randomization bias, those who got a severance package did not enroll in any training or other related labor programs, etc. – it serves to illustrate the importance of using control groups when evaluating the impact of labor programs.

example, if the mean employment rate for participants in a training program is 60% and that for non-participants is 50%, then the program impact is 10%.)

The random selection of participants is likely to lead to the absence of (or significant reduction in) selection bias among participants. However, it is often difficult to design and implement an experimental evaluation because of the following problems:

- Failure to assign randomly. This could simply be due to nepotism or could involve the exclusion of high-risk groups in order for the program administrators to show better results;
- Ethical questions about excluding some people from the intervention. This is somewhat related to the issue above. Program administrators may resist implementing the programs on the grounds that services are denied to the control group;
- Changed behavior upon learning of assignment. This could happen because individuals in an experiment know that they are part of a treatment group and act differently as a consequence; and
- Extensive data requirements. Other than being very costly, this can often be impractical as in many countries – particularly developing countries – rigorous evaluations are usually designed after a program is in place. Furthermore, there may be a significant time lag between participation in a program and follow-up surveys after the completion of the program. Many developing countries do not have the data-collection infrastructure required to follow individuals over such a lengthy time.

Econometric techniques can control for some of these problems, but they could also bias the results.

Quasi-Experimental Techniques. In these techniques, the treatment and control groups are selected after the intervention. In order to get unbiased estimates of program impact, the comparison group must be similar to the treatment group in characteristics that affect the outcomes of interest. While some of these characteristics (such as age, gender and level of education) are observable, others (such as innate ability and motivation) are not. To isolate the effect of the program, econometric techniques are used to correct for differences in the characteristics of the two groups.

Quasi-experimental evaluations are of three different types:

(i) Regression-adjusted for observables. When the observable characteristics (e.g. age, education) of the participant and the control or comparison groups differ, regression techniques can be used to compute estimates of a program impact. This is appropriate when the difference between the participant and comparison samples can be entirely explained by observable characteristics.

(ii) Regression-adjusted for observed and unobservable variables (selectivity-corrected). Simple regression techniques cannot, for obvious reasons, correct for unobservable differences between the participant and control groups. When selection into programs is not random – that is, when participation is due to both observable and unobservable characteristics – impact estimates derived from the technique in (i) above are likely to be biased. The problem is that the unobservable differences between the two groups might have caused the non-participants to have different responses to the program if they had participated. Econometric techniques have been developed to try to control for these differences (for details see Benus and Orr, O' Leary *et al.*).

(iii) Matching techniques. The control and treatment groups are likely to have different success rates in finding employment, even in the absence of ALMPs, because of differences in their observable characteristics. To control for these spurious differences, synthetic control groups are constructed. The synthetic control group, a subset of the entire control group, is composed of individuals whose observable characteristics most closely match the treatment group (there are different types matching techniques – for details see Baker).

The main appeals of quasi-experimental techniques are that they use existing data sources and are hence relatively low cost, and that these evaluations can be done at any time after the program has begun.

However, there are disadvantages. Statistical complexity is a key one: adjusting for differences in observable attributes (e.g. gender, education) is relatively straightforward but subject to specification errors; adjusting for unobservable characteristics (e.g., motivation, innate ability) requires procedures that can yield different results depending upon specification (Box 4).

Box 4: Impact Estimates For Participation In Retraining Programs, Hungary

In response to rising unemployment following the transition to a market-oriented economy, the Hungarian government instituted a wide range of labor market programs in 1990. One of these programs involved retraining. Quasi-experimental techniques were used to analyze the impact of the training for 1992 graduates of training institutions. Using different methodologies, significantly different estimates of the impact were computed.

Estimation methodology	Employment Probability (%)	Earnings Gain ($/month)
Simple difference in means	19.2*	14.9
Quasi-Experimental Techniques		
Matched Pairs	1.2	20.5
Correcting for Observables	6.3*	4.9
Correcting for Obs. and Unobservables	32.0*	na

(*–Statistically significant)

On trying different specifications, the evaluators concluded that the high estimates obtained using the correcting-for-unobservables technique were extremely sensitive to the empirical specification used. They felt that these estimates were unreliable and that the true employment impact of the program lay between the 1.2% and 6.3% generated by the matched-pairs and the correcting-for-observables techniques respectively

Relative Strengths of Techniques

It is clear from above that the absence of a control group results in the least reliable evidence of program impacts. Such techniques give no explicit estimate of what would have happened in the absence of the program, and so they provide little indication of the effects of the program. While these techniques can produce some indication of the gross outcomes of programs (e.g. number of unemployed served), policymakers should not rely on them to make comparisons across programs or decisions relating to the allocation of resources.

Experimental techniques may be the most appropriate in terms of rigor and relevance and are now more regularly implemented in many OECD countries. However, in many countries they may not be feasible owing to their high costs, excessive data requirements and the practical constraint that evaluations must be designed before the programs are underway.

Among quasi-experimental evaluations, selectivity-controlled techniques which aim to control for unobservable characteristics, may be the least appropriate. Analysis has shown that these techniques are very sensitive to the empirical specification chosen, rendering the estimates suspect.

Regression-adjusted techniques are relatively simple to perform. While they do not control for unobservable characteristics, they can be applied in cases where the treatment and control groups are roughly similar in terms of observable characteristics. Matching techniques aim to mimic experimental evaluations by removing observations in the control group that do not closely "match" with the treatment group; however these techniques are not able to control for unobserved characteristics. Still, regression-adjusted and matching techniques are possibly preferable to experimental techniques in many developing countries owing to their relatively low cost and higher feasibility.

The Importance of Costs

For the purposes of informing policy decisions, an evaluation is not complete until one considers the costs of both the ALMP and its alternatives. Cost-benefit analysis is the standard method of aggregating benefits and costs across outcome categories and across time. A program may be effective in the sense of creating benefits for participants (e.g. higher earnings and employment) but not be worthwhile if the benefits are less than

the costs involved. Unfortunately, costs appear to be the least analyzed aspect of active labor market programs.

There are two types of costs associated with programs – private and social costs. *Private costs* are those incurred by the individual. They include his/her foregone earnings while participating in the program, plus any fees or incidental expenses the individual incurs during the program. *Social costs,* on the other hand, include society-at-large's spending on the program. Hence, a societal cost computation would include the private costs as well as, for example, the rental of buildings, equipment costs and teacher salaries. In most studies that policymakers undertake, the social costs are used to evaluate cost-effectiveness.

The main steps involved in estimating costs include:

- Identifying all costs, whether or not they will be charged to the program. (For example, even if premises for a training project are provided free of charge by the government, a cost for rent should be imputed for these premises.)
- Estimating the accounting costs. This is the actual amount paid for the goods and services (e.g. salaries and benefits for administrative staff, cost of equipment and buildings).
- Including the private costs. (These include foregone earnings and any costs incurred by the individual on training.)

Cost-benefit/effectiveness evaluations may also help determine whether ALMPs reduce government spending. A program might, for example, succeed in moving people into productive employment and off unemployment benefits (a saving for the government). At the same time, the costs of the program might exceed those savings and so, on balance, actually increase government spending. A proper cost-benefit evaluation would estimate the net cost to the government.

Data Requirements

Informational constraints can play a significant role in the type of evaluation conducted. All evaluations of ALMPs require data on earnings and employment (outcome measures). For regression analysis, quasi-experimental techniques require data on socio-economic characteristics (e.g. age, education, gender, region) as well as some data on the program (e.g. length and type of training for a training program) and local labor-market characteristics (e.g. regional unemployment rates). These data would also be useful in the case of experimental evaluations if the evaluators intend to do some sub-group analysis.

Usually the data on participants is drawn through special baseline and follow-up surveys which track participants over time since their entry into the program. In the case of experimental evaluations, non-participants are also similarly tracked over time. Quasi-experimental evaluations may rely on other sources to collect information on control groups – e.g. household or labor-force surveys – though usually these will not provide the same type of information as a survey focused on particular programs will.

Several data are critical to successful evaluations (for details see O' Leary):

(a) sample selection: Participants and control-group members should have similar labor market status and eligibility for participation in programs. As the eligibility for participation in programs is usually conditioned by registration at an employment service (especially in most developed and transition economies), the register of unemployed job-seekers can be used as the sampling frame.

(b) sample size: Samples should be large enough to allow precise estimates. Larger sample sizes will permit the detection of sub-group program impacts that may be of interest to policymakers.

(c) site selection: Practical consideration should be given to the region to be canvassed. This can have significant cost implications, particularly when dealing with remote and hard to reach areas.

(d) follow up: To ensure the impact of the program is appropriately measured, it may be necessary to conduct follow-up surveys for one or two years after program completion for both treatment and control groups.

Cost data can be collected from the institutions that administer and implement labor market programs.

Who Should Conduct Evaluations?

An issue facing policymakers is whether evaluations should be conducted by government agencies or by institutions outside government. The answer to this question is critical as it will determine the development of a country's evaluation capacity. Policymakers must consider a number of factors (see Benus and Orr).

Policymakers must recognize that quantitative evaluations require highly skilled practitioners. It takes considerable expertise to develop the capacity to perform competent quantitative evaluations. In many OECD countries, these skills have been developed over the past three to four decades and the most accomplished quantitative evaluators are most likely to be found in the private sector.

Another factor to consider is the objectivity of the evaluations. As government officials are often involved in the design and implementation of ALMPs, government researchers may not be completely objective in evaluations. Political pressures to report positive results may put into question their objectivity. To reduce this type of pressure, governments sometimes establish independent units to perform evaluations. While this approach may have benefits, it does not completely eliminate the problem, especially with respect to public perceptions.

Public perception is perhaps the most difficult issue to resolve. Policy makers must recognize that the general public and the legislature may not accept the results of evaluations that are conducted by a governmental agency. This is especially true if the evaluation unit is in the ministry that is responsible for designing and implementing the program.

In many OECD countries, governments have found that it is less expensive to encourage the private sector to develop the necessary capacity to perform quantitative evaluations. In most cases, governments use this capacity for the evaluation of specific government programs. A second reason for using outside evaluators is that the results of the evaluations are likely to be objective and more readily accepted by the general public.

Irrespective of who conducts the evaluation, it is crucial that developing countries should place great emphasis on:

- training in evaluation approaches and methods;
- developing quality evaluation standards;
- strengthening monitoring systems for data on program inputs, outputs and results;
- ensuring objectives are clear, indicators are agreed upon and baselines known; and
- strengthening government's ability to disseminate the results.

Conclusions

The effectiveness of ALMPs is substantially improved if impact evaluations are rigorous and the results fed back into program design. Although carrying out rigorous evaluations may be a time-consuming and, at times, costly exercise, the long-term benefits and payoffs are substantial.

Annotated Bibliography

Baker, J. (2000). *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners.* A World Bank Publication. Aimed at providing policymakers and project managers with the tools needed to evaluate project impacts. Provides extensive case studies of a wide range of evaluations.

Benus, J. and L. Orr (2000). *Study of Alternative Quantitative Evaluation Methodologies. Working Paper.* ABT Associates, Washington D.C. Provides an overview of the importance of conducting evaluations, evaluation techniques and who should conduct evaluations.

Dar, A. and Z. Tzannatos (1999). *Active Labor Market Programs: A Review of the Evidence from Evaluations.* Social Protection Discussion Paper No. 9901. Provides a brief overview of ALMPs and evaluation techniques and presents cross-country evidence on the impacts of different ALMPs.

Grubb, W. and P. Ryan (2000). *The Roles of Evaluation for Vocational Education and Training.* ILO, Geneva. Focused on vocational training but provides an overview of evaluation techniques and methodologies.

O' Leary, C., A. Nesporova and A. Samorodov (2001). *Manual on Evaluation of Labor Market Policies in Transition Economies.* International Labour Office. Discusses various labor market programs in transition countries, evaluation methodology and how to use evaluation results.

Schmid, G., J. O' Reilly and K. Schomann (1996). *International Handbook of Labor Market Policy and Evaluation.* Edward Elgar Books. Outlines the various methodological approaches adopted in evaluation research, presents cross-country evaluation findings, and presents an insight into institutional frameworks and monitoring and evaluation systems.

Annex 1

Some Commonly Used Terms in the Impact-Evaluation Literature

Additionality: This is the net increase in jobs created. It is the total number of subsidized jobs less deadweight, substitution and displacement effects.

Deadweight Loss: Program outcomes are no different from what would have happened in the absence of the program. For example, wage subsidies place a worker in a firm that would have hired the worker in the absence of the subsidy.

Displacement Effect: This usually refers to displacement in the product market. A firm with subsidized workers increases output but displaces output among firms without subsidized workers.

Randomization Bias: This refers to bias in random-assignment experiments. In essence, this says that the behavior of individuals in an experiment will be different because of the experiment itself and not because of the goal of the experiment. Individuals in an experiment know that they are part of a treatment group and may act differently, as could individuals in the control group.

Selection Bias: Program outcomes are influenced by unobservables not controlled for in an evaluation process (e.g. individual ability). Such factors can arise as a by-product of the selection process into programs where individuals "most likely to succeed" are selected into the program.

Substitution Effect: A worker hired in a subsidized job is substituted for an unsubsidized worker who otherwise would have been hired. The net employment effect is thus zero.

Treatment and Control Group: Program beneficiaries are the "treatment" group. In a scientific evaluation, their outcomes are compared with a "control" group of non-participants.