cific case studies or anecdotes, which can give only partial information and may not be representative of overall program impacts. In this sense, well-designed and well-implemented evaluations are able to provide convincing and comprehensive evidence that can be used to inform policy decisions and shape public opinion. The summary in box 1.1 illustrates

Box 1.1: Evaluations and Political Sustainability

The Progresa/Oportunidades Conditional Cash Transfer Program in Mexico

In the 1990s, the government of Mexico launched an innovative conditional cash transfer (CCT) program called "Progresa." Its objectives were to provide poor households with short-term income support and to create incentives to investments in children's human capital, primarily by providing cash transfers to mothers in poor households conditional on their children regularly attending school and visiting a health center.

From the beginning, the government considered that it was essential to monitor and evaluate the program. The program's officials contracted a group of researchers to design an impact evaluation and build it into the program's expansion at the same time that it was rolled out successively to the participating communities.

The 2000 presidential election led to a change of the party in power. In 2001, Progresa's external evaluators presented their findings to the newly elected administration. The results of the program were impressive: they showed that the program was well targeted to the poor and had engendered promising changes in households' human capital. Schultz (2004) found that the program significantly improved school enrollment, by an average of 0.7 additional years of schooling. Gertler (2004) found that the incidence of illness in children decreased by 23 percent, while adults reported a 19 percent reduction in the number of sick or disability days. Among the nutritional outcomes, Behrman and Hoddinott (2001) found that the program reduced the probability of stunting by about 1 centimeter per year for children in the critical age range of 12 to 36 months.

These evaluation results supported a political dialogue based on evidence and contributed to the new administration's decision to continue the program. For example, the government expanded the program's reach, introducing upper-middle school scholarships and enhanced health programs for adolescents. At the same time, the results were used to modify other social assistance programs, such as the large and less well-targeted tortilla subsidy, which was scaled back.

The successful evaluation of Progresa also contributed to the rapid adoption of CCTs around the world, as well as Mexico's adoption of legislation requiring all social projects to be evaluated.

Sources: Behrman and Hoddinott 2001; Gertler 2004; Fiszbein and Schady 2009; Levy and Rodriguez 2005; Schultz 2004; Skoufias and McClafferty 2001.

how impact evaluation contributed to policy discussions around the expansion of a conditional cash transfer program in Mexico.[1] Box 1.2 illustrates how impact evaluation helped improve the allocations of the Indonesian government resources by documenting which policies were most effective in decreasing fertility rates.

Box 1.2: Evaluating to Improve Resource Allocations

Family Planning and Fertility in Indonesia

In the 1970s, Indonesia's innovative family planning efforts gained international recognition for their success in decreasing the country's fertility rates. The acclaim arose from two parallel phenomena: (1) fertility rates declined by 22 percent between 1970 and 1980, by 25 percent between 1981 and 1990, and a bit more moderately between 1991 and 1994; and (2) during the same period, the Indonesian government substantially increased resources allocated to family planning (particularly contraceptive subsidies). Given that the two things happened contemporaneously, many concluded that it was the increased investment in family planning that had led to lower fertility.

Unconvinced by the available evidence, a team of researchers tested whether family planning programs indeed lowered fertility rates. They found, contrary to what was generally believed, that family planning programs only had a moderate impact on fertility, and they argued that instead it was a change in women's status that was responsible for the decline in fertility rates. The researchers noted that before the start of the family planning program very few women of reproductive age had finished primary education. During the same period as the family planning program, however, the government undertook a large-scale education program for girls, so that by the end of the program, women entering reproductive age had benefited from that additional education. When the oil boom brought economic expansion and increased demand for labor in Indonesia, educated women's participation in the labor force increased significantly. As the value of women's time at work rose, so did the use of contraceptives. In the end, higher wages and empowerment explained 70 percent of the observed decline in fertility—more than the investment in family planning programs.

These evaluation results informed policy makers' subsequent resource allocation decisions: funding was reprogrammed away from contraception subsidies and toward programs that increased women's school enrollment. Although the ultimate goals of the two types of programs were similar, evaluation studies had shown that in the Indonesian context, lower fertility rates could be obtained more efficiently by investing in education than by investing in family planning.

Sources: Gertler and Molyneaux 1994, 2000.

received a project, program, or policy to a comparison group that did not in order to estimate the effectiveness of the program.

Beyond answering this basic evaluation question, evaluations can also be used to test the effectiveness of program implementation alternatives, that is, to answer the question, *When a program can be implemented in several ways, which one is the most effective?* In this type of evaluation, two or more approaches within a program can be compared with one another to generate evidence on which is the best alternative for reaching a particular goal. These program alternatives are often referred to as "treatment arms." For example, when the quantity of benefits a program should provide to be effective is unclear (20 hours of training or 80 hours?), impact evaluations can test the relative impact of the varying intensities of treatment (see box 1.3 for an example). Impact evaluations testing alternative program treatments normally include one treatment group for each of the treatment arms, as well as a "pure" comparison group that does not receive any program intervention. Impact evaluations can also be used to test innovations or implementation alternatives within a program. For example, a program may wish to test alternative outreach campaigns and select one group to receive a mailing campaign, while others received house-to-house visits, to assess which is most effective.

Box 1.3: Evaluating to Improve Program Design
Malnourishment and Cognitive Development in Colombia

In the early 1970s, the Human Ecology Research Station, in collaboration with the Colombian ministry of education, implemented a pilot program to address childhood malnutrition in Cali, Colombia, by providing health care and educational activities, as well as food and nutritional supplements. As part of the pilot, a team of evaluators was tasked to determine (1) how long such a program should last to reduce malnutrition among preschool children from low-income families and (2) whether the interventions could also lead to improvements in cognitive development.

The program was eventually made available to all eligible families, but during the pilot, the evaluators were able to compare similar groups of children who received different treatment durations. The evaluators first used a screening process to identify a target group of 333 malnourished children. These children were then classified into 20 sectors by neighborhood, and each sector was randomly assigned to one of four treatment groups. The groups differed only in the sequence in which they started the treatment and, hence, in the amount of time that they spent in the program. Group 4 started the earliest and was exposed to the treatment for the longest period, followed by groups 3, 2, and then 1. The treatment itself consisted of 6 hours of health care and

(continued)

Box 1.3 *continued*

educational activities per day, plus additional food and nutritional supplements. At regular intervals over the course of the program, the evaluators used cognitive tests to track the progress of children in all four groups.

The evaluators found that the children who were in the program for the longest time demonstrated the greatest gains in cognitive improvement. On the Stanford-Binet intelligence test, which estimates mental age minus chronological age, group 4 children averaged –5 months, and group 1 children averaged –15 months.

This example illustrates how program implementers and policy makers are able to use evaluations of multiple treatment arms to determine the most effective program alternative.

Source: McKay et al. 1978.

Deciding Whether to Evaluate

Not all programs warrant an impact evaluation. Impact evaluations can be costly, and your evaluation budget should be used strategically. If you are starting, or thinking about expanding, a new program and wondering whether to go ahead with an impact evaluation, asking a few basic questions will help with the decision.

The first question to ask would be, *What are the stakes of this program?* The answer to that question will depend on both the budget that is involved and the number of people who are, or will eventually be, affected by the program. Hence, the next questions, *Does, or will, the program require a large portion of the available budget?* and, *Does, or will, the program affect a large number of people?* If the program does not require a budget or only affects a few people, it may not be worth evaluating. For example, for a program that provides counseling to hospital patients using volunteers, the budget involved and number of people affected may not justify an impact evaluation. By contrast, a pay reform for teachers that will eventually affect all primary teachers in the country would be a program with much higher stakes.

If you determine that the stakes are high, then the next question is whether any evidence exists to show that the program works. In particular, do you know how big the program's impact would be? Is the available evidence from a similar country with similar circumstances? If no evidence is available about the potential of the type of program being contemplated, you may want to start out with a pilot that incorporates an impact evaluation. By contrast, if evidence is available from similar circumstances, the

discuss in detail how to collect cost data or conduct cost-benefit analysis.[2] However, it is critically important that impact evaluation be complemented with information on the cost of the project, program, or policy being evaluated. Once impact and cost information is available for a variety of programs, cost-effectiveness analysis can identify which investments yield the highest rate of return and allow policy makers to make informed decisions on which intervention to invest in. Box 1.4 illustrates how impact evaluations can be used to identify the most cost-effective programs and improve resource allocation.

Box 1.4: Evaluating Cost-Effectiveness

Comparing Strategies to Increase School Attendance in Kenya

By evaluating a number of programs in a similar setting, it is possible to compare the relative cost-effectiveness of different approaches to improving outcomes such as school attendance. In Kenya, the nongovernmental organization International Child Support Africa (ICS Africa) implemented a series of education interventions that included treatment against intestinal worms, provision of free school uniforms, and provision of school meals. Each of the interventions was subjected to a randomized evaluation and cost-benefit analysis, and comparison among them provides interesting insights on how to increase school attendance.

A program that provided medication against intestinal worms to schoolchildren increased attendance by approximately 0.14 years per treated child, at an estimated cost of $0.49 per child. This amounts to about $3.50 per additional year of school participation, including the externalities experienced by children and adults not in the schools but in the communities that benefit from the reduced transmission of worms.

A second intervention, the Child Sponsorship Program, reduced the cost of school attendance by providing school uniforms to pupils in seven randomly selected schools. Dropout rates fell dramatically in treatment schools, and after 5 years the program was estimated to increase years in school by an average of 17 percent. However, even under the most optimistic assumptions, the cost of increasing school attendance using the school uniform program was estimated to be approximately $99 per additional year of school attendance.

Finally, a program that provided free breakfasts to children in 25 randomly selected preschools led to a 30 percent increase in attendance in treatment schools, at an estimated cost of $36 per additional year of schooling. Test scores also increased by about 0.4 standard deviations, provided the teacher was well trained prior to the program.

Although similar interventions may have different target outcomes, such as the health effects of deworming or educational achievement in addition to increased participation, comparing a number of evaluations conducted in the same context can reveal which programs achieved the desired goals at the lowest cost.

Sources: Kremer and Miguel 2004; Kremer, Moulin, and Namunyu 2003; Poverty Action Lab 2005; Vermeersch and Kremer 2005.

- What are the nature and scope of the problem? Where is it located, whom does it affect, how many are affected, and how does the problem affect them?
- What is it about the problem or its effects that justifies new, expanded, or modified social programs?
- What feasible interventions are likely to significantly ameliorate the problem?
- What are the appropriate target populations for intervention?
- Is a particular intervention reaching its target population?
- Is the intervention being implemented well? Are the intended services being provided?
- Is the intervention effective in attaining the desired goals or benefits?
- Is the program cost reasonable in relation to its effectiveness and benefits?

Answers to such questions are necessary for local or specialized programs, such as job training in a small town, a new mathematics curriculum for elementary schools, or the outpatient services of a community mental health clinic, as well as for broad national or state programs in such areas as health care, welfare, and educational reform. Providing those answers is the work of persons in the program evaluation field. valuators use social research methods to study, appraise, and help improve social programs, including the soundness of the programs' diagnoses of the social problems they address, the way the programs are conceptualized and implemented, the outcomes they achieve, and their efficiency. (Exhibit 1-A conveys the views of one feisty senator about the need for evaluation evidence on the effectiveness of programs.)

EXHIBIT 1-A

A Veteran Policymaker Wants to See the Evaluation Results

But all the while we were taking on this large—and, as we can now say, hugely successful—effort [deficit reduction], we were constantly besieged by administration officials wanting us to *add* money for this social program or that social program…. *My* favorite in this miscellany was something called "family preservation," yet another categorical aid program (there were a dozen in place already) which amounted to a dollop of social services and a press release for some subcommittee chairman. The program was to cost $930 million over five years, starting at $60 million in fiscal year 1994. For three decades I had been watching families come apart in our society; now I was being told by seemingly everyone on the new team that one more program would do the trick…. At the risk of indiscretion, let me include in the record at this point a letter I wrote on July 28, 1993, to Dr. Laura D' Andrea Tyson, then the distinguished chairman of the Council of Economic Advisors, regarding the Family Preservation program:

Dear Dr. Tyson:

You will recall that last Thursday when you so kindly joined us at a meeting of the Democratic Policy Committee you and I discussed the President's family preservation proposal. You indicated how much he supports the measure. I assured you I, too, support it, but went on to ask what evidence was there that it would have any effect. You assured me there were such data. Just for fun, I asked for two citations.

The next day we received a fax from Sharon Glied of your staff with a number of citations and a paper, "Evaluating the Results," that appears to have been written by Frank Farrow of the Center for the Study of Social Policy here in Washington and Harold Richman at the Chapin Hall Center at the University of Chicago. The paper is quite direct: "Solid proof that family preservation services can affect a state's overall placement rates is still lacking."

Just yesterday, the same Chapin Hall Center released an "Evaluation of the Illinois Family First Placement Prevention Program: Final Report." This was a large scale study of the Illinois Family First initiative authorized by the Illinois Family Preservation Act of 1987. It was "designed to test effects of this program on out-of-home placement and other outcomes, such as subsequent child maltreatment." Data on case and service characteristics were provided by Family First caseworkers on approximately 4,500 cases: approximately 1,600 families participated in the randomized experiment. The findings are clear enough.

Overall, the Family First placement prevention program results in a slight increase in placement rates (when data from all experimental sites are combined). This effect disappears once case and site variations are taken into account. In other words, there are either negative effects or no effects.

This is nothing new. Here is Peter Rossi's conclusion in his 1992 paper, "Assessing Family Preservation Programs." Evaluations conducted to date "do not form a sufficient basis upon which to firmly decide whether family preservation programs are either effective or not." May I say to you that there is nothing in the least surprising in either of these findings? From the mid-60s on this has been the repeated, I almost want to say consistent, pattern of evaluation studies. Either few effects or negative effects. Thus the negative income tax experiments of the 1970s appeared to produce an increase in family breakup.

This pattern of "counterintuitive" findings first appeared in the '60s. Greeley and Rossi, some of my work, and Coleman's. To this day I cannot decide whether we are dealing here with an artifact of methodology or a much larger and more intractable fact of social programs. In any event, by 1978 we had Rossi's Iron Law. To wit: "If there is any empirical law that is emerging from the past decade of widespread evaluation activity, it is that the expected value for any measured effect of a social program is zero."

I write you at such length for what I believe to be an important purpose. In the last six months I have been repeatedly impressed by the number of members of the Clinton administration who have assured me with great vigor that something or other is known in an area of social policy which, to the best of my understanding, is not known at all. This seems to me perilous. It is quite possible to live with uncertainty, with the possibility, even the likelihood that one is wrong. But beware of certainty where none exists. Ideological certainty easily degenerates into an insistence upon ignorance.

The great strength of political conservatives at this time (and for a generation) is that they are open to the thought that matters are complex. Liberals got into a reflexive pattern of denying this. I had hoped

twelve years in the wilderness might have changed this; it may be it has only reinforced it. If this is so, current revival of liberalism will be brief and inconsequential.

Respectfully,

Senator Daniel Patrick Moynihan

SOURCE: Adapted, with permission, from D. P. Moynihan, *Miles to Go: A Personal History of Social Policy* (Cambridge, MA: Harvard University Press, 1996), pp. 47-49.

Although this text emphasizes the evaluation of social programs, especially human service programs, program evaluation is not restricted to that arena. The broad scope of program evaluation can be seen in the evaluations of the U.S. General Accounting Office (GAO), which have covered the procurement and testing of military hardware, quality control for drinking water, the maintenance of major highways, the use of hormones to stimulate growth in beef cattle, and other organized activities far afield from human services.

Indeed, the techniques described in this text are useful in virtually all spheres of activity in which issues are raised about the effectiveness of organized social action. For example, the mass communication and advertising industries use essentially the same approaches in developing media programs and marketing products. Commercial and industrial corporations evaluate the procedures they use in selecting, training, and promoting employees and organizing their workforces. Political candidates develop their campaigns by evaluating the voter appeal of different strategies. Consumer products are tested for performance, durability, and safety. Administrators in both the public and private sectors often assess the managerial, fiscal, and personnel practices of their organizations. This list of examples could be extended indefinitely.

These various applications of evaluation are distinguished primarily by the nature and goals of the endeavors being evaluated. In this text, we have chosen to emphasize the evaluation of social programs—programs designed to benefit the human condition—rather than efforts that have such purposes as increasing profits or amassing influence and power. This choice stems from a desire to concentrate on a particularly significant and active area of evaluation as well as from a practical need to limit the scope of the book. Note that throughout this book we use the terms *evaluation, program evaluation,* and *evaluation research* interchangeably.

To illustrate the evaluation of social programs more concretely, we offer below examples of social programs that have been evaluated under the sponsorship of local, state, and federal government agencies, international organizations, private foundations and philanthropies, and both nonprofit and for-profit associations and corporations.

- In several major cities in the United States, a large private foundation provided funding to establish community health centers in low-income areas. The centers were intended as an alternative way for residents to obtain ambulatory patient care that they could otherwise obtain only from hospital outpatient clinics and emergency rooms at great public cost. It was further hoped that by improving access to such care, the clinics might increase timely treatment and thus reduce the need for lengthy and