

Reflexive controls	Measures of the outcome variable taken on participating targets before interventions and used as control observations.
Pre-post design	A reflexive control design in which only one or a few before-intervention and after-intervention measures are taken.
Shadow controls	Expert and participant judgments used to estimate net impact.
Time-series analyses	Reflexive designs that rely on relatively long series of repeated measurements of the outcome variable taken before and after an intervention.

ASSESSMENT OF FULL-COVERAGE PROGRAMS

In this chapter, we consider designs for assessing the impact of full-coverage programs. Of course, many programs intended to reach all targets fail to do so because of either faulty delivery or insufficient interest by eligible targets. In designing evaluations, it may be possible to treat these as partial-coverage programs and configure a comparison group. Often, however, there are not enough unserved targets available to construct comparison groups; under these circumstances, the most common option is to assess changes on outcome measures that occur between a point before targets participate in a program and some point afterward. The term reflexive controls is used to describe impact assessments of this sort where targets serve as their own controls. Reflexive control designs generally provide only weak evaluations of impact because, in contrast to random and quasi-experimental evaluations, it is not generally possible to take certain important confounding effects into account. When it is possible to obtain a large number of pre- and post-program measurements on outcome variables, sophisticated time-series analyses can be undertaken, which usually allow for firmer assessment of outcomes. An alternative to reflexive controls is the use of shadow controls. Shadow controls basically consist of knowledgeable experts, administrators, or targets themselves, who judge program outcomes in the light of their experience or opinion regarding what might have occurred without the intervention. Shadow control evaluations rarely produce completely convincing findings about impact.

Evaluators of fully or almost fully saturated programs, ones in which virtually all eligible targets participate in the program, encounter circumstances that make it very difficult to undertake impact assessments that yield credible results. In the absence of a comparison group, it is virtually impossible to take all the confounding effects into account that are likely to be influential and to produce convincing

estimates of program effects. With the possible exception of some applications of time-series analyses in which there are a large number of measurements on outcome variables, evaluators undertaking assessments of full-coverage programs should expect the results to be vulnerable to criticism.

At the same time, it is essential to evaluate a myriad of full-coverage programs. Many, if

not most, government-sponsored social programs are mandated to cover all targets in the population; indeed, often it is the right of all eligible targets not only to participate but to receive the same intensity or array of services. Obvious examples at the national level are Medicare and Social Security for elderly persons and, at the local level, public safety and schooling. In these instances, there is little possibility of forming a reasonable comparison group, which necessarily means that the opportunity for firm estimates of impact is limited. It is evident that in comparison with randomized and quasi-experimental studies, the level of confidence in the effectiveness estimates of full-coverage programs generally is low. However, it is not possible to simply avoid evaluating such programs because decisionmakers and the public persistently and appropriately raise questions about their effectiveness.

This chapter discusses two types of designs applicable to impact assessments of full-coverage programs: those using *reflexive controls* and judgmental assessments using what we have termed *shadow controls*. In reflexive control studies, the estimation of net impact comes entirely from information on the targets at two or more points in time, at least one of which is before exposure to the program. Shadow controls are based on the assumption that it is possible for some persons to estimate the impact of programs by comparing program outcomes to their conjectures about what could be expected to occur without the program. Almost always, judgmental approaches are suspect for estimating impact.

Because of the limitations of the designs discussed in this chapter, evaluators are wise to consider transforming a study of a supposedly full-coverage program into one that considers the relative effectiveness of variations in the program, perhaps consisting of differences in

the intensity or "dosage" provided different targets or even in qualitative differences in the way a program is implemented from jurisdiction to jurisdiction. Under such circumstances, it may be possible to use quasi-experiments that approximate those discussed in Chapter 9. We discuss this possibility first.

NONUNIFORM FULL-COVERAGE PROGRAMS

There are many cases where supposedly uniform programs do not actually deliver the same intervention at the same strength and intensity to all targets. Nonuniform full-coverage programs, in which implementation varies significantly, can be subjected to impact evaluations that assess the differential effects of variations in the program. However, it should be noted that such assessments provide estimates of the effects of the more effective variations relative to the less effective ones, and not of the effects of the program relative to no program.

The government subsidies and tax credits for child care assessed by Fuller et al. (Exhibit 10-A) provide a good example of a program that varies in level of activity from area to area. By examining the differing levels of these support programs across 36 states and making use of sophisticated statistical controls, the evaluators were able to estimate their effects on five indicators of the quality of child care providers despite the fact that no state was without such programs.

A classic cross-sectional study of a full-coverage program is the Coleman report (Coleman et al., 1966). This study assessed the impact of differences among schools in staffing levels, finances, student composition, and physical

EXHIBIT 10-A Using State-Level Policy Variation to Assess the Impact of Subsidies and Tax Credits on the Quality of Child Care Centers

The 1990 federal budget bill created the first nationwide child care program with \$3 billion for tax credit, voucher, and state block grants in addition to the \$1.1 billion spent directly by state governments for subsidies or vouchers. But does this government investment result in higher levels of child care quality? Fuller, Raudenbush, Wei, and Holloway set out to answer this question by examining the relationship between variation in the type and amount of government support and the quality of child care providers.

They drew their data from a nationally representative sample of 2,089 child care centers across 36 states that was collected for a study conducted by Mathematica Policy Research; state-level information about subsidies, regulations, and staff training requirements; demographic indicators aggregated to the state level from the Bureau of the Census and the Children's Defense Fund; and Internal Revenue Service summaries of state-by-state utilization of the Child and Dependent Care Tax Credit and the Earned Income Tax Credit.

SOURCE: Adapted from Bruce Fuller, Stephen W. Raudenbush, Li-Ming Wei, and Susan D. Holloway, "Can Government Raise Child-Care Quality? The Influence of Family Demand, Poverty, and Policy," *Educational Evaluation and Policy Analysis*, 1993, 15(3):255-278.

Using hierarchical linear modeling, the researchers examined the association of three types of state child care policies and two indicators of the inflow of tax credit subsidies to the states with five indicators of the quality of child care centers. Control variables were used to take account of the influence of family demand for child care; state-level wealth, maternal employment, and poverty rates; organizational type and structure of the child care centers, and the ethnicity of the participating children.

Government subsidies to the child care centers were found to be related to staff with higher qualifications, larger teacher salaries, more formalized instructional programs, and more frequent parent participation but not to child-staff ratios. However, level of state regulation and utilization of the Child and Dependent Care Tax Credit were not independently related to the quality of the child care centers.

plants on student learning. Coleman's original finding was that differences in these variables were not related very strongly to student achievement. Holding such student background variables constant, he found that students achieved no more in schools spending a great deal per capita than in schools spending considerably less. Similar findings held for student-to-teacher ratios, the adequacy of physical plants, and the training of teachers.

The Coleman report was not universally acclaimed as definitive, however. Many educators and researchers disputed the findings and produced a spate of reanalyses that tested alternative statistical models on the same data. This case again illustrates the vulnerability of one-shot surveys to criticisms on grounds of specification errors.

A final example involves the use of a simple before-and-after reflexive design (Exhibit 10-B).

EXHIBIT 10-B Estimating the Effects of Zoning Regulations on Housing and Population

Using local municipal records, Shlay and Rossi ascertained the zoning regulations in force in 1960 for each census tract in a sample from the Chicago metropolitan area. A set of indexes was constructed for each tract reflecting the extent to which the zoning regulations restricted residential use of tract land, ranging from the most exclusionary use pattern, in which only single-family homes on large tracts were permitted, to the least exclusionary usage, in which any type of land use, including industrial and commercial use, was permitted.

Using 1960 and 1970 census reports for housing and population characteristics, the researchers conducted a regression analysis that predicted the 1970 density of housing units in suburban tracts on the basis of 1960 density and

the zoning index. Note that this analysis is reflexive in that the 1960 housing density is statistically controlled by including that variable in the regression equation. Thus, what is being studied is the difference between actual 1970 density and that expected on the basis of the 1960 density measures. The other independent variables are all measures of the zoning regulations governing land use in the tracts as of 1960.

The results showed that the more restrictively a tract was zoned, the less its density grew in the period between 1960 and 1970. In short, exclusionary zoning restricted the growth in housing density in suburban tracts over and above what would be expected on the basis of their density in 1960.

SOURCE: Adapted, with permission, from A. Shlay and P. H. Rossi, "Keeping Up the Neighborhood: Estimating Net Effects of Zoning," *American Sociological Review*, October 1981, 46:703-719.

Shlay and Rossi (1981) obtained data on a sample of census tracts in the Chicago metropolitan area to assess the effects of zoning regulations on population and housing growth in the tracts. This impact assessment took advantage of the considerable variation in zoning regulations from tract to tract, using the 1960 and 1970 decennial censuses as before and after measures. Using regression analysis, Shlay and Rossi arrived at estimates of the effects of zoning restrictions on the growth of housing density.

Note that the analysis shown in Exhibit 10-B depends heavily on the existence of variation from census tract to census tract in the 1960 zoning regulations. Hence, each tract

serves as its own control in predicting growth in the period between censuses, and tracts are contrasted according to the amount of restriction placed on land uses in each. The researchers took into account the *maturational trends* in tract growth, such as age-related changes in individuals, by estimating such trends for the entire set of tracts and by considering zoning the cause of deviations from such maturational trends, as represented by the predicted 1970 values of housing and population stocks.

For many programs, however, there may not be significant variation to permit this approach to impact assessment, or evaluation sponsors may not be interested in the effects of

program variations. In such cases, it usually becomes necessary to resort to one of the approaches described in the remainder of this chapter.

REFLEXIVE CONTROLS

When reflexive controls are used, the presumption must be made that no changes in the targets on the outcome variables have occurred in the time between observations other than those induced by the intervention. Under this presumption, any difference between preintervention status on those variables and postintervention status are deemed net intervention effects. For example, suppose that pensioners from a large corporation previously received their checks in the mail but now have them automatically deposited in their bank accounts. Comparison of complaints about late or missing payments before and after this procedure was implemented could be construed as evidence of impact, provided that it was plausible that the rate of burglaries from mailboxes, the level of postal service, and so on had not also changed.

The strongest reflexive control designs are time series consisting of a large number of observations over the time period spanning the intervention. For example, suppose that instead of just a pre- and postmeasure of pensioners' complaints, we had monthly information for, say, two years before and one year after the change in payment procedures. In this case, our degree of certainty about net effects would be higher because we would have more information on which to base our estimates about what would have happened had there been no change in the mode of check delivery. A second procedure often used is to disaggregate the outcome data by various characteristics of the targets.

For example, examining time-series data about pensioners' complaints regarding receipt of checks in high- and low-crime areas and in rural and urban areas would provide additional insight into the impact of the change in procedure.

In general, the use of reflexive controls is not recommended in circumstances in which comparison or control groups are possible. However, there may be formative evaluation situations in which reflexive controls are appropriate. For example, if a very large difference is essential for an innovative program to be widely implemented, it may be sensible to test it out with a pretest-posttest design that does not include a comparison group. Although only knowledge of the gross effect would emerge from the evaluation, if the gross effect is well below the needed change, there is no basis for continuing plans for implementation. Of course, if a sufficient gross effect is found, it becomes important to undertake a more refined evaluation to have net effect information.

Thus, reflexive controls may be applied usefully to partial-coverage programs as an economical first step, especially if there is little reason to believe that targets' scores on outcome measures would have changed without the intervention. For instance, the impact of a schoolwide nutrition education program consisting of a three-week set of lectures might be evaluated by testing students' knowledge of nutrition before and after participation. The use of reflexive controls in this circumstance is likely to provide a reasonable estimate of the effects of the program because knowledge of nutrition is unlikely to change spontaneously over such a short period, although addition of a properly constructed comparison group would increase confidence.

The impact assessment formula for reflexive designs is as follows:

$$\text{Net effect} = \left[\begin{array}{c} \text{Outcomes for} \\ \text{participants} \\ \text{after} \\ \text{intervention} \end{array} \right] - \left[\begin{array}{c} \text{Outcomes for} \\ \text{participants} \\ \text{before} \\ \text{intervention} \end{array} \right] \pm \left[\begin{array}{c} \text{Effects} \\ \text{of other} \\ \text{processes at} \\ \text{work during} \\ \text{intervention} \end{array} \right] + \left[\begin{array}{c} \text{Design} \\ \text{effects and} \\ \text{stochastic} \\ \text{error} \end{array} \right]$$

Note that the critical term in the assessment formula represents the effects of other processes at work during intervention. This refers to all those influences that act to bring about change on the outcome variables independently from the intervention. All reflexive designs are vulnerable to such influences because of their lack of control for this class of effects and the general difficulty of developing good estimates of them without a comparison group.

There are several variations of reflexive designs that depend principally on how many measures are taken before and after the program in question has gone into effect. We will examine each of these variations in turn.

Simple Pre-Post Studies

A simple pre-post (or before and after) study is one in which one set of measurements is taken on targets before program participation and a second set is taken on the same targets after sufficiently long participation for effects to be expected. Impact is estimated by comparing the two sets of measurements.

As we have noted, the main drawback to this design is that the differences between before and after measures cannot be confidently ascribed to program effects. All the processes at work in the intervening period may affect these differences. For example, it might be tempting

to assess the effectiveness of Medicare by comparing the health statuses of persons before they became eligible with the same measures taken after a few years of participation in Medicare. Such comparisons would be badly misleading for a variety of reasons. In the first place, the maturational effects of aging generally lead to poorer health on their own. In addition, retirement may change the life circumstances of the participants so drastically as to affect their health. Income changes also occur at the same time, and there are also the effects of the deaths of spouses, friends, and so on.

The particular dangers of relying on pre-post reflexive designs for behavior that is age related are demonstrated in the following illustration. Consider the evaluation of a program directed toward women of childbearing age that is designed to lower fertility. The evaluation attempts to show that participation in the program lowers the probability of conceiving, as compared to the ten-year period preceding program participation. Such comparisons would be quite misleading because fertility behavior at one point in time is not independent of prior fertility behavior. Some women will have completed their fertility at the point of program participation and would not have had any more children in any event. Others may be just beginning their families; because they have not had previous children, the birth of any child will appear as a failure of the program. In short, the processes at work producing fertility vary strongly with age, and hence a reflexive design would not yield good estimates of net program effects.

Sometimes time-related changes are more subtle. For example, reflexive controls are likely to be questionable in studies of the impact of job training programs. One of the main reasons people choose to enter such programs is that

they are unemployed and are experiencing difficulties obtaining employment. Hence, at the time of entry into the program, most participants have no or very depressed income and some of them are likely to locate jobs irrespective of participation in the program. A job training program will thus appear to be successful automatically if only reflexive controls are used in its evaluation.

A second problem with reflexive controls arises out of potential changes in secular trends between the two time periods involved. If, in a program to increase crop yields, preprogram observations of farmers are made during a period of depressed yields, a comparison with yields during a subsequent period of more normal growing conditions would be misleading. Similarly, a program to reduce crime will appear more effective if it coincides with, say, efforts to increase policing. Confounding factors can also skew an assessment in the other direction: An employment training program will appear ineffective if it is accompanied by a prolonged period of rising unemployment and depressed economic conditions.

A third problem results from possible interfering events between the two points of data collection. An interfering event, as defined previously, is an unusual, one-time occurrence that affects outcome measures. Examples include natural occurrences (e.g., storms), political events, and outbreaks of other social problems. Any intervening event that might affect output measures could interfere with the validity of reflexive control designs for estimating program effects.

In general, then, simple pre-post reflexive designs usually provide findings that have a low degree of credibility. This is particularly the case when the time elapsed between the two measurements is appreciable—say, a year or more—because over time it becomes more and

more likely that some process occurring during the time period may obscure the effects of the program, whether by enhancing or by diminishing them. The simple pre-post design, therefore, is appropriate mainly for short-term impact assessments of full-coverage programs attempting to affect conditions that are unlikely to change much on their own (Exhibit 10-C provides an example of such a situation).

Complex Repeated Measures Reflexive Designs

Panel studies that involve repeated measurements on the same group over a period of time can often be used to produce estimates of net intervention effects that have a fair degree of credibility. The reason is that the participation of targets will often vary over time, and the "dips and peaks" can be tied in with the shifts in outcome measures. Although such opportunities are not present in all instances, we urge evaluators to take advantage of them when possible. Fundamentally, this is the same advice we offered earlier in this chapter about taking advantage of differences in intensity or amount of intervention to convert reflexive control evaluations to comparison group quasi-experiments. Even if there is no comparison group, if variations in participation can be tied to changes in the outcome measures, the credibility of the findings is increased.

Exhibit 10-D describes an elaborate attempt by Milavsky and colleagues (1982) to estimate the impact of viewing violence on television on children's subsequent aggressive behaviors. Some exposure to television is almost universal among young children. However, the assessment was made possible by the considerable variation from child to child in the amount of viewing and the contents of the programs they viewed.

EXHIBIT 10-C A Convincing Pre-Post Outcome Design for a Program to Reduce Residential Lead Levels in Low-Income Housing

The toxic effects of lead are especially harmful to children and can impede their behavioral development, reduce their intelligence, cause hearing loss, and interfere with important biological functions. Poor children are at disproportionate risk for lead poisoning because the homes available to low-income tenants are generally older homes, which are more likely to be painted with lead paint and to be located near other sources of lead contamination. Interior lead paint deteriorates to produce microscopic quantities of lead that children may ingest through hand-to-mouth activity. Moreover, blown or tracked-in dust may be contaminated by deteriorating exterior lead paint or roadside soil containing a cumulation of lead from the leaded gasoline used prior to 1980.

To reduce lead dust levels in low-income urban housing, the Community Lead Education and Reduction Corps (CLEARCorps) was initiated in Baltimore as a joint public-private effort. CLEARCorps members clean, repair, and make homes lead safe, educate residents on lead-poisoning prevention techniques, and encourage the residents to maintain low levels of lead dust through specialized cleaning efforts. To determine the extent to which CLEARCorps was

successful in reducing the lead dust levels in treated urban housing units, CLEARCorps members collected lead dust wipe samples immediately before, immediately after, and six months following their lead hazard control efforts. In each of 43 treated houses, four samples were collected from each of four locations—floors, window sills, window wells, and carpets—and sent to laboratories for analysis.

Statistically significant differences were found between pre and post lead dust levels for floors, window sills, and window wells. At the six-month follow-up, further significant declines were found for floors and window wells, with a marginally significant decrease for window sills.

Since no control group was used, it is possible that factors other than the CLEARCorps program contributed to the decline in lead dust levels found in the evaluation. Other than relevant, but modest, seasonal effects relating to the follow-up period and the small possibility that another intervention program treated these same households, for which no evidence was available, there are few plausible alternative explanations for the decline. The evaluators concluded, therefore, that the CLEARCorps program was effective in reducing residential lead levels.

SOURCE: Adapted from Jonathan P. Duckart, "An Evaluation of the Baltimore Community Lead Education and Reduction Corps (CLEARCorps) Program," *Evaluation Review*, 1998, 22(3):373-402.

The advantage of panel studies is that the measures of the intervention and outcomes (e.g., TV viewing and aggressiveness, respectively) are related to each other through time lags and not as cross-sectional correlations. Thus, aggressiveness at Time 2 is examined as

a function of viewing patterns measured at Time 1. Panel studies are especially appropriate for impact assessments of full-coverage programs whose dosage varies over individuals and over time. In the case of TV viewing, all the children participated in the sense that virtually

EXHIBIT 10-D Measuring the Effects of TV Violence on Children's Aggressive Behavior

In an attempt to provide rigorous answers to public concern over whether the viewing of TV programs depicting violence and aggression affect children's aggressive behavior, the National Broadcasting Company (NBC) sponsored an elaborate panel study of young children in which aggressiveness and TV viewing were measured repeatedly over several years.

In the main substudy, samples of elementary school classes, Grades 2 through 6, drawn from Fort Worth and Minneapolis schools, formed the base for a six-wave panel study, in which 400 male children in 59 classes were interviewed six times in the period 1970 to 1973. (Additional substudies were conducted with female elementary school children and with samples of high school students in the same cities.) At each interview wave, the children in the classes were asked to rate each other on aggressiveness using questionnaires that included such items as "Who is likely to punch and kick another child?" The questionnaires also picked up information about the socioeconomic background of the children.

SOURCE: Adapted from J. R. Milavsky, H. H. Stipp, R. C. Kessler, and W. S. Rubens, *Television and Aggression: A Panel Study* (New York: Academic Press, 1982).

all viewed some TV. Nevertheless, some children viewed more programs containing violence than others and some watched more such programs at some times than at other times. Self-selection was to some degree controlled by statistically controlling the initial level of aggressiveness of the children under study.

It should be noted that the researchers in this study considered using randomized experiments to estimate the effects of viewing violent programs on subsequent aggressiveness but re-

In addition, at every interview, the children were each asked to check those programs they had watched recently on lists of programs shown locally. The programs previously had been rated by media experts according to the amount of violence depicted in them. To check the accuracy of recall, several nonexistent programs were placed on the checklists. Additional interviews were conducted with the children's teachers and parents.

The analyses undertaken related the viewing of violence on TV at one interview time with rated aggressive behavior at subsequent interview times, controlling statistically for the initial level of the children's aggressiveness. The results estimated the additional amount of aggressiveness that resulted from high levels of exposure to violence on TV programs. While the direction of effects indicated a small increment in aggressiveness associated with high levels of viewing of TV violence, that increment was not statistically significant.

jected that design as introducing an artificiality that would undermine the generalizability of their findings. It would be difficult, if not impossible, to recruit schoolchildren for experimentation, randomly allocate them to experimental and control groups, and then somehow prevent the controls from viewing any programs that contain aggressive or violent behavior. An experiment along those lines might be conducted for a very short period of time, on the order of a few days, but would be extremely

difficult to carry out over the length of time needed to show the expected effects. In other words, the researchers opted for a less than optimal design of lower validity for estimating effects but higher generalizability.

Time-Series Evaluations

The term *panel study* usually designates research using a relatively modest number of repeated measurements taken on the same group of respondents to study the effects of a program or other types of change. In impact assessment, panel studies may not have measures of preprogram participation or status on outcome variables (as in the example in Exhibit 10-D) but only measures of exposure to program differentials and subsequent outcome status.

Repeated measures time-series designs, more generally, may involve many measurements that cover the periods before the program has been put into place as well as afterward and may not include the same respondents at each time of measurement. Extensive time-series data of some sort are often available to track changes in some of the conditions that social programs address and that, therefore, may be relevant to impact assessment. Many continuing databases compile periodic information related to phenomena of major public concern (e.g., fertility, mortality, and crime) or administrative concern (e.g., proportions of college students dropping out at the end of their first year). Such time series can provide relatively firm bases on which to build estimates of the net effects of full-coverage programs.

Most existing times series involve aggregated data such as averages or rates computed for one or more political jurisdictions. For ex-

ample, the Department of Labor maintains an excellent time series that has tracked unemployment rates monthly for the country as a whole and major regions since 1948. Monthly rates are also available for major population subgroups—by sex, age, race, level of educational attainment, and so on.

When a relatively long time series of preintervention observations exists, it is often possible to model long-standing trends in the target group, projecting those trends through and beyond the time of the intervention and observing whether the postintervention period shows significant deviations from the projections. The use of such general time-trend modeling procedures as ARIMA (auto regressive integrated moving average; see Hamilton, 1994; McCleary and Hay, 1980) can identify the best-fitting statistical models by taking into account long-term secular trends and seasonal variations. They also allow for the degree to which any value or score obtained at one point in time is necessarily related to previous ones (technically referred to as *autocorrelation*). The procedures involved are technical and require a fairly high level of statistical sophistication.

Exhibit 10-E illustrates the use of time-series data for assessing the impact of raising the legal drinking age on alcohol-related traffic accidents. This evaluation is made possible by the existence of relatively long series of measures on the outcome variable (more than 200). The analysis uses information collected over the eight to ten years prior to the policy changes of interest to establish the expected trends for alcohol-related accident rates for different age groups legally entitled to drink. Comparison of the age-stratified rates experienced after the drinking ages were raised with the expected rates based on the prior trends provides a measure of net impact.

EXHIBIT 10-E Estimating the Effects of Raising the Drinking Age From Time-Series Data

During the early 1980s, many states raised the minimum drinking age from 18 to 21, especially after passage of the Federal Uniform Drinking Age Act of 1984, which reduced highway construction funds to states that maintained a drinking age less than 21. The general reason for this was the widespread perception that lower drinking ages had led to dramatic increases in the rate of alcohol-related traffic accidents among teenagers. Assessing the impact of raising the drinking age, however, is complicated by downward trends in accidents stemming from the introduction of new automobile safety factors and increased public awareness of the dangers of drinking and driving.

Wisconsin raised its drinking age to 19 in 1984, then to 21 in 1986. To assess the impact of these changes, David Figlio examined an 18-year time series of monthly observations on alcohol-related traffic accidents, stratified by age, that was available from the Wisconsin Department of Transportation for the period from 1976 to 1993. Statistical time-series models were fit to the data for 18-year-olds (who could legally drink prior to

1984), for 19- and 20-year-olds (who could legally drink prior to 1986), and for over-21-year-olds (who could legally drink over the whole time period). The outcome variable in these analyses was the rate of alcohol-related crashes per thousand and licensed drivers in the respective age group.

The results showed that, for 18-year-olds, raising the minimum drinking age to 19 reduced the alcohol-related crashes by an estimated 26% from the prior average of 2.2 per month per 1,000 drivers. For 19- and 20-year-olds, raising the minimum drinking age to 21 reduced the monthly crash rate by an estimated 19% from an average of 1.8 per month per 1,000 drivers. By comparison, the estimated effect of the legal changes for the 21-and-over group was only 2.5% and statistically nonsignificant.

The evaluator's conclusion was that the imposition of increased minimum drinking ages in Wisconsin had immediate and conclusive effects on the number of teenagers involved in alcohol-related crashes, resulting in substantially fewer than the prelegislation trends would have generated.

SOURCE: Adapted, by permission, from David N. Figlio, "The Effect of Drinking Age Laws and Alcohol-Related Crashes: Time-Series Evidence From Wisconsin," *Journal of Policy Analysis and Management*, 1995, 14(4):555-566. Copyright © 1995, John Wiley & Sons, Inc.

As noted earlier, the units of analysis in time-series data relevant to social programs are usually highly aggregated. Exhibit 10-E deals essentially with one case, the state of Wisconsin, where accident measures are constructed by aggregating the pertinent data over the entire state and expressing them as accident rates per 1,000 licensed drivers. The statistical models

developed to fit such data are vulnerable to specification error just like all the other such models we have discussed. For example, if there were significant influences on the alcohol-related accident rates in Wisconsin at certain times that were not represented in the trend lines estimated by the model, then the results of the analysis would not be valid.

Simple graphic methods of examining time-series data before and after an intervention can provide crude but useful clues to impact. Indeed, if the confounding influences on an intervention are known and there is considerable certainty that their effects are minimal, simple examination of a time-series plot may identify obvious program effects. Exhibit 10-F presents the primary data for one of the classic applications of time series in program evaluation: the British Breathalyzer crackdown (Ross, Campbell, and Glass, 1970). The graph in that exhibit shows the auto accident rates in Great Britain before and after the enactment and enforcement of drastically changed penalties for driving while under the influence of alcohol. The accompanying chart indicates that the legislation had a discernible impact: Accidents declined after it went into effect and the decline was especially dramatic for accidents occurring over the weekend when we would expect higher levels of alcohol consumption. Although the effects are rather evident in the graph, it is wise to confirm them with statistical analysis; the reductions in accidents visible in Exhibit 10-F are, in fact, statistically significant.

Time-series approaches are not necessarily restricted to single cases, however. When time series exist for interventions at different times and in different places, more complex analyses can and should be undertaken. Parker and Rebhun (1995), for instance, examined the relationship of changes in state minimum age of purchase laws for alcohol with homicide rates with time series covering 1976-1983 for each of the 50 states plus the District of Columbia. Parker and Rebhun used a pooled cross-section time-series analysis with a dummy code (0 or 1) to identify the years before and after the drinking age was raised. Other variables in the

model included alcohol consumption (beer sales in barrels per capita), infant mortality (as a poverty index), an index of inequality, racial composition, region, and total state population. This model was applied to homicide rates for different age groups, and raising the minimum age of purchase law was found to be significantly related to reductions in homicide for victims in the age 21-24 category.

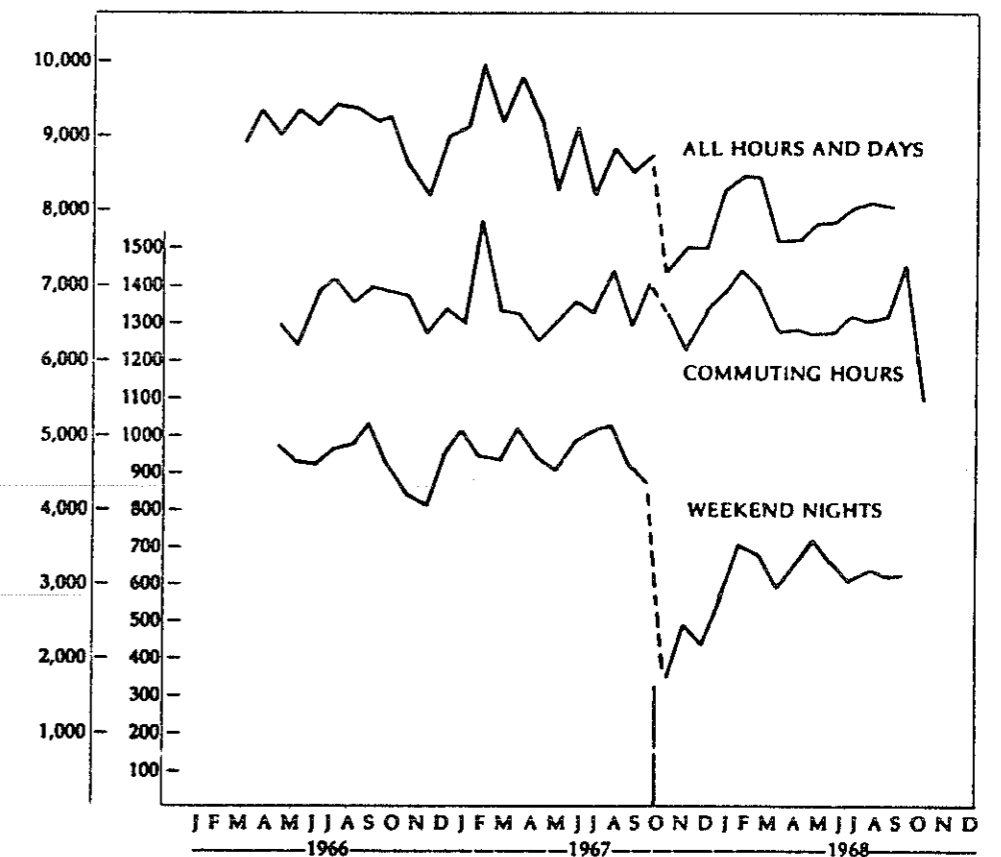
Although the time-series analyses discussed above all use aggregated data, the general logic of time-series analyses can also be applied to disaggregated data, as in the analysis of interventions administered to single cases or small groups of persons whose behavior is measured a number of times before, after, and perhaps, during program participation. Therapists, for example, have used time-series designs to assess the impact of treatments on individual clients. Thus, a child's performance on some achievement test may be measured periodically before and after a new teaching method is used with the child, or an adult's drinking behavior may be measured before and after therapy for alcohol abuse. The logic of time-series analyses remains the same when applied to a single case, although the statistical methods applied are different because the issues of long-term trends and seasonality usually are not as serious for individual cases (Kazdin, 1982).

In general, when appropriate data are available, time-series analyses are relatively strong designs for estimating the effects of instituting uniform, full-coverage programs or the effects of making changes in such programs. We recommend them for circumstances in which appropriate statistical series exist, with the caveat that such analyses are vulnerable to specification errors.

EXHIBIT 10-F An Analysis of the Impact of Compulsory Breathalyzer Tests on Traffic Accidents

In 1967 the British government enacted a new policy that allowed police to give Breathalyzer tests at the scenes of accidents. The test measured the presence of alcohol in the blood of suspects. At the same time, heavier penalties were instituted for drunken-driving convictions. Considerable publicity was given to the provisions of the new law, which went into effect in October 1967.

The chart below plots vehicular accident rates by various periods of the week before and after the new legislation went into effect. Visual inspection of the chart clearly indicates that a decline in accidents occurred after the legislation, which affected most times of the week but had especially dramatic effects for weekend periods. Statistical tests verified that these declines are greater than could be expected from the chance component of these data.



SOURCE: Adapted from H. L. Ross, D. T. Campbell, and G. V. Glass, "Determining the Social Effects of a Legal Reform: The British Breathalyzer Crackdown of 1967," *American Behavioral Scientist*, March/April 1970, 13:494-509.

SHADOW CONTROLS

When one of the more scientifically credible approaches to impact assessment is either not possible or not cost-effective, it is tempting to seek some less demanding but still reasonable alternative. The judgment of experts is one such alternative. There are persons with expertise in various human service areas on whose judgment it may be possible to rely when constructing estimates of whether a given gross outcome is sufficient indication that there has been an appreciable net impact. In addition, it is also possible to solicit judgments from program administrators or turn to the participants themselves to obtain assessments of how a program has affected them. We have termed the judgments of experts, program administrators, and participants *shadow controls*, a name chosen to reflect their role as a benchmark for comparison as well as their usual lack of a substantial evidential basis.

Shadow controls are used for a variety of reasons. Sometimes they are seen as an option to reflexive control designs for full-coverage programs, other times they are used because of their minimal costs, and still other times they are simply the traditional means used in a particular program area. In general, these judgmental assessments—even when highly trained, expert judges are used—rarely yield convincing estimates of net effects. At worst, they are highly unreliable and misestimate even gross program effects (see Exhibit 10-G). We include them in this book because of the frequency with which they are used rather than because we advocate using them.

Despite the fact that shadow controls ordinarily must be viewed with considerable skepticism, there are circumstances in which they may be justified. Although it may be sensible

to spend a lot of time and money on the evaluation of prospective social programs that, if enacted, would be costly or might produce harmful unintended effects, there are many programs that are not that important. For example, it probably was justified to spend several hundred million dollars on the randomized experiments that tested the effectiveness of income maintenance programs, because a national income maintenance program would be very costly. As a matter of public policy, it may not be worth expending much money or effort on the evaluation of a program that is small (and intended to remain so) and in which the intervention is not a significant cost. In such cases, shadow controls may provide a rough estimate of impact that is sufficient for the information needs of the relevant decision-makers.

Another circumstance in which the use of shadow controls may be justified is the case of an extraordinarily and obviously successful program. For instance, suppose we found that a two-month-long vocational training program to produce drivers of heavy-duty trucks has enabled 90% of the participants (selected from among persons without such skills) to qualify for appropriate driver's licenses. Such a finding suggests that the program has been quite successful in reaching its goal of imparting vocational skills. We can make this judgment because it seems highly unlikely that so large a proportion of any group of previously unskilled persons who wanted to become truck drivers would be able to qualify for the licenses in a two-month period on their own. In this case, the shadow control estimate is based on generally held knowledge about motor vehicle licensing; if the evaluator wanted substantiation of the "shadow control" percentage, all he or she would have to do is call up the state motor vehicle licensing agency.

EXHIBIT 10-G Evaluating Family Preservation Programs: How Shadow Controls Can Be Misleading

Family preservation programs have been initiated in many areas to provide intensive casework to families of abused or neglected children judged "at imminent risk" of being placed in foster care. These programs provide caseworkers who meet with families frequently, as much as several days per week, for periods of six to eight weeks. The function of the caseworkers is to help the families handle crises, impart proper parenting behavior, and teach strategies for coping with disciplinary difficulties.

The initial evaluations of family preservation programs were based on shadow controls. Every child in the program who, after participating, was not placed in foster care was counted as a "success" based on the assumption that they would have gone into foster care without the

program. Using this estimate of outcome without service, the agencies running family preservation programs claimed startlingly high success rates, varying from 75% to 90%.

When randomized experiments were conducted, however, a very different story emerged. Those families and children who were placed in control groups and received ordinary child protective services—often meaning little or no services—were found to be no more nor no less likely to be placed in foster care than those who participated in a family preservation program. In one large experiment in Illinois, for instance, 16% of the children in control group families and 16% of those in experimental families were placed in foster care.

SOURCE: Adapted from Peter H. Rossi, "Assessing Family Preservation Programs," *Children and Youth Services Review*, 1992, 14(1):77-98.

Of course, the obverse outcomes could also lead to firm judgments. If all of the program's participants failed the license examination, that finding would be fairly clear evidence of the program's failure. Even this seemingly straightforward example, however, illustrates the risk of using shadow controls. If the evaluator did not check with the licensing agency, he or she might not know that the layperson's idea about failure rates may not be correct. It may well be, for example, that almost all applicants fail the first try and that the crucial percentage is the proportion who pass on the second attempt.

In reality, of course, the observed outcome would probably be more ambiguous—say, only

30% passing the first time. This more typical finding raises the question of whether a comparable group receiving no training would have done as well. The shadow control information available is only a rough estimate of the expected proportion who would pass without special training, and in any case there is no way of removing confounding effects from the gross outcome. In short, without a comparison group that is the equivalent of the participating targets, there is no way to estimate the net effects of the program with a high degree of certainty. Simply knowing the proportion in a program who "succeeded" and comparing that proportion to what might be expected does not provide

an estimate of net program effect—hence the warning that relying on shadow controls for impact assessment is always risky.

Connoisseurial Assessments

Expert or "connoisseurial" judgments are the most common form of shadow controls. For instance, persons familiar with adult vocational education and the typical outcomes of intervention programs in that field might be asked to draw on their background to judge whether a 30% outcome in our truck driving example represents a success given the nature of the targets. Clearly, the usefulness and validity of such judgments, and hence the worth of an evaluation using them, depend heavily on the judges' expertise and knowledge of the program area.

Whenever shadow controls of any kind are to be used, their basis should be made explicit. Thus, the reasoning and assumptions on which expert judgments are made need to be described as completely as possible. If an expert makes a judgment based on his or her own direct experiences, the extent of those experiences ought to be revealed. When possible, explicit references to other evaluation studies should be given so that others can check whether the circumstances of the other interventions are comparable to those of the one under judgment. (See Nevo, 1989, for a set of rules for using experts and their judgments in evaluations.)

The actual procedures employed by experts to arrive at shadow controls vary considerably. Often, one or more well-known experts in a relevant field are hired as consultants to visit the site of a program, examine its workings closely, and write a report summarizing their assessment. Visiting experts may examine program records, observe the program in opera-

tion, conduct interviews with current and former participants, and talk to program managers, staff, and other officials. In short, all the means of informal social research may be employed.

Often the shadow control judgment of a connoisseur is a conclusion or construction based on the expert's understanding of the processes involved. The worth of the resulting assessments varies accordingly. Thus, to an expert in criminology, it may "stand to reason" that a given intervention concerning the rehabilitation of ex-prisoners will be effective, because it closely follows the leading theories in the field. However, whereas an industrial engineer's judgment concerning the effectiveness of a production process may be quite enough to justify action, the judgment of a criminologist about rehabilitation simply does not command the same standing. Unfortunately, the very reason for employing rigorous impact assessment designs in the area of social programs is that the state of knowledge in the appropriate fields is inadequate. Although it "stands to reason" that many programs will succeed, they often do not pass the more rigorous tests of the better impact assessment designs.

The worth of an expert's judgmental assessment depends on several factors. First, one must consider the general state of knowledge in the substantive fields relevant to the program. In a field where knowledge of how to achieve a particular outcome is quite advanced, an expert's appraisal may be very accurate. If little is known about an area (e.g., how to rehabilitate criminals), an expert's judgment of a particular program's effectiveness may be of little more merit than that of any informed person.

Second, one must consider how well grounded the expert is in the substantive field. An expert should be knowledgeable about the

area in question and should have demonstrated that knowledge in actual accomplishments.

Third, experts should be familiar with the findings of evaluations of similar programs, especially systematic ones. For example, an expert asked to judge whether a community intervention center for released prisoners helps ex-felons obtain employment should be familiar with the many studies on the employment rates for ex-felons during the months after release. Similarly, knowledge that few studies show much influence of classroom size on achievement, whereas studies do show strong positive effects on teachers' satisfaction ratings, should make one skeptical that a program based largely on reducing class size is likely to do much for student achievement (although it may please the teachers).

Finally, because experts often rely heavily on "guided" visits to program sites, and often have more contact with program staff than with program participants, it is essential to be cautious about judgments based heavily on the reports of program staff. After all, it is only natural for a program administrator to attempt to present the program in the best possible light. Thus, one can expect the state of a program at the time of an announced visit to be better than at other periods in ways ranging from the neatness of the headquarters to possibly well-rehearsed laudatory statements from participants.

A skillful expert will base his or her judgments on data obtained from many sources. At a minimum, experts should consider the following:

1. *Administrative records.* Experts should collect information from administrative records (or have such tabulations made) on such topics as

- a. Size of the program
- b. Type of participants recruited
- c. Attrition experience with participants
- d. Postprogram experiences of participants
- e. Program costs per participant who completes the program
- f. Participant changes relevant to program goals, assessed by before-and-after measures

2. *Observations of program operation.* In programs that call for active work with participants (e.g., household visits, classroom sessions, media presentations), the work should be directly observed by the visiting experts.

3. *Interviews with participants.* Informal interviews with participants and/or former participants, at least some of which are spontaneous, can take up such issues as

- a. Recruitment of participants
- b. Motivation of participants
- c. Participant satisfaction with the program
- d. Participant progress toward attaining program goals

4. *Interviews with stakeholders and informants on the program's context.* Informal interviews with local officials, administrators of competing programs, administrators of important local institutions (e.g., school superintendents, members of the clergy, police chiefs), and local powerful individuals or representatives of local powerful institutions (e.g., large landlords, bankers, political officials) should cover the following topics:

- a. Worth of the program
- b. The extent to which the program is viewed as a help or a threat to the community
- c. Interest in continuing the program when the demonstration period is over

Despite the weaknesses of connoisseurial assessments, there are circumstances in which this approach may be the only one that is feasible. This may be the case, for instance, with some full-coverage, constant-level interventions of long standing. Also, the urgency of the need for impact assessment may force evaluators to rely on expert judgments, or a lack of resources may prohibit them from mounting an impact assessment using control groups or a cross-sectional approach. Moreover, other controls may be feasible in principle but prohibitively tedious to put into practice.

When the funds allocated to evaluation are inadequate for a full-scale impact assessment, connoisseurial judgments may be resorted to on the grounds that some assessment is better than none at all. This may be particularly the case when the program is on a small scale and it would appear incongruous if the impact assessment cost a large fraction of the program costs. Under these circumstances, the "good enough" rule propounded in Chapter 7 may be invoked. Of course, a responsible connoisseurial evaluation will warn consumers of the judgmental basis for the findings.

Program Administrator Judgments

Program administrators are routinely asked to assess their progress toward fulfilling program goals. In most cases, it is doubtful that much reliance can be placed on such assess-

ments, for fairly obvious reasons. First, undertaking a judgmental impact assessment is a difficult task under the best of circumstances. But the use of program administrator judgments is especially weak in light of the difficulty they will naturally have in adopting the attitude of skepticism toward their own work that is necessary for making hard judgments. A properly conducted impact assessment takes as its guiding hypothesis that the program has no effects, a stance that runs exactly counter to the principle that should guide the administration of a program—namely, that the intervention does have important effects on participants. To expect ordinary mortals to hold both hypotheses simultaneously is unrealistic. Furthermore, program administrators who have day-to-day responsibilities for the conduct of a program, and who often lack appropriate technical qualifications for making assessments of impact, often simply cannot devote a great deal of time and care to such assessments. In addition, there is an understandable tendency for administrators to want to put their program in the best of all possible lights, a motivation that may lead them to downplay or actively suppress negative information on effectiveness.

About the best one can expect from an administrator's judgmental assessment is reasonably accurate descriptive statements about operational procedures. One is entitled to expect that administrators will provide reliable statistical, descriptive statements about a program based on a well-kept set of administrative records. The kinds of records necessary have been described earlier in this chapter and in Chapter 6. Exhibit 10-H lists additional administrative records that can be useful for impact assessments. Clearly, all of these records are not appropriate to all programs, and this list should therefore be regarded as suggested rather than essential records.

EXHIBIT 10-H Administrative Records Useful for Project Description and as Aids in Impact Assessment

1. Participant records
 - a. Socioeconomic data on participants: age, sex, location, household composition, income, occupational data
 - b. Critical dates: date of entry into program, attendance record, date of leaving program
 - c. Service records: exposure of participants to the program, services provided, and so forth
 - d. Follow-up data: addresses of participants, including future addresses and contacts to aid in follow-up beyond participation
 - e. Critical event records: records of meetings with participants, important events in participants' lives (e.g., births, deaths, residential shifts, job changes)
2. Program records
 - a. Critical events in the program history: dates of start-up for important components of the program, encounters with helpful or hostile officials, periods when the program was not operational
 - b. Program personnel: biographical data on program personnel, shifts in personnel, record of personnel training
 - c. Changes in program implementation: problems encountered, changes instituted in program operations (including dates)
3. Financial records

No attempt to describe such records will be made here since one can assume that the fiscal procedures typically required by program sponsors will be employed. The main issue to be emphasized is that financial records should be kept in a way that will facilitate cost-effectiveness or cost-benefit analyses, as described in Chapter 11.

Participant Judgments

Because the participants in social programs are the recipients of the services, evaluators might be tempted to rely on their accounts of how well they were served by the program as indications of net impact. Although participants can tell us many useful things, it is overly

optimistic to expect that they can provide a direct assessment of program impact. The problem is that it is difficult for any individual to realistically assess what would have happened to him or her if some specific event had not occurred. Most people simply do not have the varied experiences or psychological distance from their own circumstances to be able

EXHIBIT 10-I The Quality of Day Care From the Perspectives of Parents

Parents tend to judge the quality of child care by different criteria than do child development specialists. One study found that parents who used better quality day care (according to experts) paid, on average, no more than parents using poorer quality services. This may indicate that parents do not value the characteristics of "quality" as highly as other characteristics of the services. Another study, for instance, provides evidence that parents may be more concerned about the location, hours, and dependability of day care arrangements than they are about aspects of quality considered important by child development professionals.

Parents appear to judge the quality of child care according to (a) whether it offers a safe and

healthy environment—many parents express considerable concern about potential child abuse; (b) whether the environment promotes learning—a concern that is especially prevalent among parents of older children; and (c) convenience, including location within a 10- to 15-minute radius of home or work and hours of operation that mesh with the mother's work schedule (e.g., accommodating shift work, overtime, and other special needs).

Parents generally apply these criteria on the basis of limited knowledge about the range of their child care options. Most providers do not advertise their services, and most users do not look at alternative arrangements before placing their child in care.

SOURCE: Adapted from Ellen Kisker and Rebecca Maynard, "Quality, Cost and Parental Choice of Child Care," in *The Economics of Child Care*, ed. David M. Blau (New York: Russell Sage, 1991).

to construct a valid image of what their condition would be had they not participated in the program. Note that this is not a view of human beings as naive or deficient, but a recognition that assessing net impact is a comparative task and that most persons lack the breadth of experience or vantage point needed to make such comparisons for their own outcomes.

Participants' ratings of satisfaction with a program or with services, however, are informative and important in their own right. In the first place, some programs stipulate participant satisfaction as one of their goals and work to rid their procedures of the "bugs" that irritate participants. Retirement benefit programs, for instance, attempt to deliver retirement income

in a way that is most satisfying to beneficiaries, including automatic bank deposits or special pick-up provisions. Public service programs may be particularly concerned with client satisfaction as an index of service-unit functioning.

As an illustration, Exhibit 10-I provides a summary of studies conducted to find out what parents considered to be quality day care arrangements for their children. Note that the perspectives of parents and those of child development experts appear to be somewhat at variance. Participant assessments of programs thus offer useful information even if they cannot replace carefully designed impact assessments.

SUMMARY

- ✖ The impact of full-coverage programs is difficult to assess with confidence because the nature of such programs precludes the use of comparison groups. For programs with nonuniform coverage, evaluators sometimes are able to take advantage of variations in the intensity, type, or amount of service to approximate quasi-experimental impact assessment designs.
- ✖ Full-coverage programs generally are evaluated by using reflexive controls to compare preprogram and postprogram outcome measurements. Designs range from simple before-and-after comparisons, with only one measurement before and one after program participation, to time series involving multiple measurements before and after the intervention is in place. Time-series designs are much better than simple pre-post designs for estimating net effects. In evaluations with only two measurements, it is almost impossible to differentiate net from gross effects.
- ✖ An alternative to reflexive designs is the use of shadow controls. Shadow control evaluations make use of judges to estimate impact. These judges may be experts in the program area, program administrators, or program participants themselves. Although such evaluations are commonplace, they are not recommended because it is difficult to obtain valid estimates of net program effects from them. Shadow controls can be used with relative confidence only when program effects are readily apparent and there is good reason to believe that most of the gross effects can be presumed to be net effects of the program.

KEY CONCEPTS FOR CHAPTER 11

Costs	Inputs, both direct and indirect, required to produce an intervention.
Benefits	Net program outcomes, usually translated into monetary terms. Benefits may include both direct and indirect effects.
Net benefits	The total discounted benefits minus the total discounted costs. Also called net rate of return.
Cost-effectiveness	The efficacy of a program in achieving given intervention outcomes in relation to the program costs.
Cost-benefit analysis	Analytical procedure for determining the economic efficiency of a program, expressed as the relationship between costs and outcomes, usually measured in monetary terms.
Benefits-to-costs ratio	The total discounted benefits divided by the total discounted costs.
Ex ante efficiency analysis	An efficiency analysis undertaken prior to program implementation, usually as part of program planning, to estimate net outcomes in relation to costs.
Ex post efficiency analysis	An efficiency analysis undertaken subsequent to knowing a program's net outcome effects.
Accounting perspectives	Perspectives underlying decisions on which categories of goods and services to include as costs or benefits in an analysis.
Shadow prices	Imputed or estimated costs of goods and services not valued accurately in the marketplace. Shadow prices also are used when market prices are inappropriate due to regulation or externalities. Also known as accounting prices.
Opportunity costs	The value of opportunities forgone because of an intervention program.
Externalities	Effects of a program that impose costs on persons or groups who are not targets.
Distributional effects	Effects of programs that result in a redistribution of resources in the general population.
Discounting	The treatment of time in valuing costs and benefits, that is, the adjustment of costs and benefits to their present values, requiring a choice of discount rate and time frame.
Internal rate of return	The calculated value for the discount rate necessary for total discounted program benefits to equal total discounted program costs.

CHAPTER 11

MEASURING EFFICIENCY

Knowledge of the extent to which programs have been implemented successfully and the degree to which they have the desired outcomes is indispensable to program managers, stakeholders, and policymakers. In almost all cases, however, it is just as critical to be informed about how program outcomes compare to their costs. In fact, whether it is accomplished impressionistically, as in most everyday life decisions, or by formal procedures, comparison of the costs and benefits of social programs is one of the most important considerations in deciding whether to expand, continue, or terminate them.

Efficiency assessments—cost-benefit and cost-effectiveness analyses—provide a frame of reference for relating costs to program results. In addition to providing information for making decisions on the allocation of resources, they are often useful in gaining the support of planning groups and political constituencies who determine the fate of social intervention efforts.

The procedures employed in both types of analyses are often highly technical, and their applications will be described only briefly in this chapter. However, because the issue of the cost or effort required to achieve a given magnitude of desired change is implicit in all impact evaluations, all program evaluators must understand the ideas embodied in efficiency analyses, even if the technical procedures are beyond their skills.

- Policymakers must decide how to allocate funding among a variety of educational programs, ranging from basic primary educational classes for young children to vocational training efforts for adults. All have been shown to have substantial net impact in completed evaluations. How should the available educational resources be allocated?
- A government agency is reviewing national disease control programs currently in operation. If additional funds are to be allocated to disease control, which programs would show the biggest payoffs per dollar of expenditure?
- Evaluations in the criminal justice field have established the effects of various alternative programs aimed at reducing recidivism. Which program is most cost-effective to the criminal justice system? Given the policy choices, how would altering the current pattern of expenditures maximize the efficiency of correctional alternatives?
- Members of a private funding group are debating whether to promote a program of low-interest loans for home construction or to initiate work skills training for married women to increase family income. How should they decide?

These are examples of common resource allocation dilemmas faced by planners, funding groups, and policymakers everywhere. Again and again, decisionmakers must choose how to allocate scarce resources to put them to optimal use. Consider even the fortunate case in which pilot projects of several programs have shown them all to be effective in producing the desired net impacts. The decision of which to fund on a larger scale must take into account the relations between costs and outcomes in each program. Although other factors, including political and value considerations, come into play, the preferred program often is the one that produces the most impact on the most targets for a given level of expenditure. This simple principle is the foundation of cost-benefit and cost-effectiveness analyses, techniques that provide systematic approaches to resource allocation analysis.

Both cost-benefit and cost-effectiveness analyses are means of judging the efficiency of programs. As we will elaborate, the difference between the two types of analyses is the way in which the outcomes of a program are expressed. In *cost-benefit* analyses, the outcomes of programs are expressed in monetary terms; in *cost-effectiveness* analyses, outcomes are expressed in substantive terms. For example, a cost-benefit analysis of a program to reduce cigarette smoking would focus on the difference between the dollars expended on the anti-smoking program and the dollar savings from reduced medical care for smoking-related diseases, days lost from work, and so on. A cost-effectiveness analysis of the same program would estimate the dollars that had to be expended to convert each smoker into a non-smoker. (Later in this chapter we discuss the basis for deciding whether to undertake a cost-benefit or cost-effectiveness analysis.)

The basic procedures and concepts underlying resource allocation analysis stem from work undertaken in the 1930s to establish decision-making criteria for public investment activities. Early applications in the United States were to water resource development; in England, to transportation investments. After World War II, organizations such as the World Bank stimulated the application of cost-benefit analysis to both specific project activities and national programs in less-developed as well as industrialized countries. Perhaps the greatest stimulus to systematic application of cost-benefit analysis to governmental programs was the Planning, Programming, and Budgeting System (PPBS) implemented in the 1960s, an extension of the systems analysis approach then being applied in the Department of Defense. (For a review of how efficiency analyses have been applied in the federal government over the years, see Nelson, 1987.)

Cost-benefit and cost-effectiveness analyses in the social program area have their analogue in the world of business, where costs are constantly compared with income. For instance, a computer company may be concerned with the relationship of costs to income for making microcomputers compatible with those of a major competitor. Or a small restaurant owner might be concerned with whether to provide dinner music or promote her lunchtime specials to increase profits.

The idea of judging the utility of social intervention efforts in terms of their efficiency (profitability, in business terms) has gained widespread acceptance. However, the question of "correct" procedures for actually conducting cost-benefit and cost-effectiveness analyses of social programs remains an area of considerable controversy (Eddy, 1992; Zerbe, 1998). As we will discuss, this controversy is related to a

combination of unfamiliarity with the analytical procedures employed, reluctance to impose monetary values on many social program outcomes, and an unwillingness to forsake initiatives that have been held in esteem for extended periods of time. Evaluators undertaking cost-benefit or cost-effectiveness analyses of social interventions must be aware of the particular issues involved in applying efficiency analyses to their specific field, as well as the limitations that characterize the use of cost-benefit and cost-effectiveness analyses in general. (For comprehensive discussions of efficiency assessment procedures, see Gramblin, 1990; Nas, 1996; Yates, 1996.)

KEY CONCEPTS IN EFFICIENCY ANALYSIS

Cost-benefit and cost-effectiveness analyses can be viewed both as conceptual perspectives and as sophisticated technical procedures. From a conceptual point of view, perhaps the greatest value of efficiency analysis is that it forces us to think in a disciplined fashion about both costs and benefits. In the case of virtually all social programs, identifying and comparing the actual or anticipated costs with the known or expected benefits can prove invaluable. Most other types of evaluation focus mainly on the benefits. Furthermore, efficiency analyses provide a comparative perspective on the relative utility of interventions. Judgments of the comparative utility of different initiatives are unavoidable, since social programs, almost without exception, are conducted under resource constraints. Almost invariably, maintaining continuing support depends on convincing policymakers and funders that the "bottom

line" (i.e., dollar benefits or the equivalent) justifies the program.

An interesting illustration of decision making along these lines is a report of a large bank's support of a day care center for its employees (see Exhibit 11-A). As the report documents, despite the difficulties of undertaking efficiency analyses, and even when they are somewhat crudely done, they can provide evidence supporting the implementation of company-supported social programs. The article from which the excerpt in Exhibit 11-A is taken also discusses preventive health programs, day care centers, and lunchtime educational programs established by various businesses. In each case, knowing the bottom line in terms of the cost of benefits was the basis of the company's decisions.

In spite of their value, however, it bears emphasis that in many evaluations formal, complete efficiency analyses are either impractical or unwise for several reasons. First, the required technical procedures may be beyond the resources of the evaluation project; may call for methodological sophistication not available to the project's staff; or may be unnecessary, given either very minimal or extremely high efficacy of the intervention. Second, political or moral controversies may result from placing economic values on particular input or outcome measures, controversies that could obscure the relevance and minimize the potential utility of an otherwise useful and rigorous evaluation. Third, expressing the results of evaluation studies in efficiency terms may require selectively taking different costs and outcomes into account, depending on the perspectives and values of sponsors, stakeholders, targets, and evaluators themselves (what are referred to as *accounting perspectives*). The dependence of results on the accounting per-

EXHIBIT 11-A Cost Savings From a Bank's Child Care Facilities

In January 1987, Union Bank opened a new profit center in Los Angeles. This one, however, doesn't lend money. It doesn't manage money. It takes care of children.

The profit center is a day-care facility at the bank's Monterey Park operations center. Union Bank provided the facility with a \$105,000 subsidy last year. In return, it saved the bank as much as \$232,000. There is, of course, nothing extraordinary about a day-care center. What is extraordinary is the \$232,000. That number is part of a growing body of research that tries to tell companies what they are getting—on the bottom line—for the dollars they invest in such benefits and policies as day-care assistance, wellness plans, maternity leaves, and flexible work schedules.

The Union Bank study, designed to cover many questions left out of other evaluations, offers one of the more revealing glimpses of the savings from corporate day-care centers. For one thing, the study was begun a year before the center opened, giving researchers more control over the comparison statistics. Union Bank approved spending \$430,000 to build its day-care center only after seeing the savings projections.

Using data provided by the bank's human resource department, Sandra Burud, a child-care

consultant in Pasadena, California, compared absenteeism, turnover, and maternity leave time the first year of operation and the year before. She looked at the results for 87 users of the center, a control group of 105 employees with children of similar ages who used other day-care options, and employees as a whole.

Her conclusion: The day-care center saves the bank \$138,000 to \$232,000 a year—numbers she calls "very conservative." Ms. Burud says savings on turnover total \$63,000 to \$157,000, based mostly on the fact that turnover among center users was 2.2 percent compared with 9.5 percent in the control group and 18 percent throughout the bank.

She also counted \$35,000 in savings on lost days' work. Users of the center were absent an average of 1.7 fewer days than the control group, and their maternity leaves were 1.2 weeks shorter than for other employees. Ms. Burud also added a bonus of \$40,000 in free publicity, based on estimates of media coverage of the center.

Despite the complexities of measurement, she says, the study succeeds in contradicting the "simplistic view of child care. This isn't a touchy-feely kind of program. It's as much a management tool as it is an employee benefit."

SOURCE: J. Solomon, "Companies Try Measuring Cost Savings From New Types of Corporate Benefits," *Wall Street Journal*, December 29, 1988, p. B1. Reprinted by permission of The Wall Street Journal, Dow Jones & Company, Inc. All rights reserved worldwide.

spective employed may be difficult for at least some of the stakeholders to comprehend, again obscuring the relevance and utility of evaluations.

Furthermore, efficiency analysis may be heavily dependent on untested assumptions or

the requisite data for undertaking cost-benefit or cost-effectiveness calculations may not be fully available. Even the strongest advocates of efficiency analyses acknowledge that there often is no single "right" analysis. Moreover, in some applications, the results may show unac-

ceptable levels of sensitivity to reasonable variations in the analytic and conceptual models used and their underlying assumptions.

Although we want to emphasize that the results of all cost-benefit and cost-effectiveness analyses should be treated with caution, and sometimes with a fair degree of skepticism, such analyses can provide a reproducible and rational way of estimating the efficiency of programs. Even strong advocates of efficiency analyses rarely argue that such studies should be the sole determinant of decisions about programs. Nonetheless, they are a valuable input into the complex mosaic from which decisions emerge.

Timing of Efficiency Analyses

The employment of cost-benefit and cost-effectiveness techniques is appropriate at all phases of program efforts. However, efficiency analyses are most commonly undertaken either during the planning and design phase of an initiative or after an innovative or markedly modified program has been in place for a time and there is interest in making it permanent or possibly expanding it.

In the planning and design phases, *ex ante* efficiency analyses may be undertaken on the basis of a program's anticipated costs and outcomes. Such analyses, of course, must presume a given magnitude of positive net impact even if this value is only a conjecture. Likewise, the costs of providing and delivering the intervention must be estimated by one means or another. In some cases, estimates of both the inputs and the magnitude of impact can be made with considerable confidence, either because there has been a pilot program (or a similar program in another location) or because the program is fairly simple in its implementation. Nevertheless, since *ex ante* analyses in

whole or in part are not based on empirical information, they run the risk of seriously under- or overestimating net benefits. Indeed, the issue of the accuracy of the estimates of both inputs and outputs is one of the controversial areas in *ex ante* analyses.

Ex ante cost-benefit analyses are most important for those programs that will be difficult to abandon once they have been put into place or that require extensive commitments in funding and time to be realized. For example, the decision to increase recreational facilities by putting in new jetties along the New Jersey ocean shore would be difficult to overturn once the jetties had been constructed; thus, there is a need to estimate the costs and outcomes of such a program compared with other ways of increasing recreational opportunities, or to judge the wisdom of increasing recreational opportunities compared with the costs and outcomes of allocating the resources to another social program area.

Thus, when extensive resource commitments would be required by program sponsors to initiate and maintain a program, decisions are preceded in many cases by *ex ante* cost-benefit analyses. Exhibit 11-B illustrates such a situation with regard to the testing of health care workers for HIV. Even though the possibility of, say, a surgeon or dentist transmitting HIV/AIDS to a patient is a matter of serious consequences and concern, testing and regulating the vast number of health care workers in this country for HIV is likely to be quite expensive. Before embarking on such a program, it is wise to develop some estimate, even if crude, of how expensive it is likely to be in relation to the number of patient infections averted. The analysis summarized in Exhibit 11-B showed that under most risk scenarios any reasonable policy option would likely be quite expensive. Moreover, there was considerable uncertainty

EXHIBIT 11-B Ex-Ante Analysis of the Cost-Effectiveness of HIV Testing for Health Care Workers

A study by Phillips and others in 1994 examined the cost-effectiveness of alternative policies for HIV testing of health care workers, including physicians, surgeons, and dentists. The policy options considered were (a) mandatory and (b) voluntary testing, and for those who test positive, (a) exclusion from patient care, (b) restriction of practice, or (c) a requirement that patients be informed of their HIV status.

The derivation of costs in this study was based on data obtained from reviewing the pertinent literature and consulting with experts. The cost estimates included three components: (a) counseling and testing costs, (b) additional treatment costs because of early detection of HIV-positive cases, and (c) medical care costs averted per patient infection averted. Costs were estimated by subtracting (c) from (a) + (b).

Analyzing all options under high, medium, and low HIV prevalence and transmission risk scenarios, the study concluded that one-time mandatory testing with mandatory restriction of practice for a health care worker found HIV positive was more cost-effective than the other

options. While showing the lowest cost of the policies considered, that option nonetheless was estimated to cost \$291,000 per infection averted for surgeons and \$500,000 for dentists. Given these high costs and the political difficulties associated with adopting and implementing mandatory restrictions on practice, this was not considered a viable policy option.

The analysts also found that the cost-effectiveness estimates were highly sensitive to variations in prevalence and transmission risk and to the different patterns of practice for physicians in contrast to dentists. The incremental cost per infection averted ranged from \$447 million for dentists under low prevalence/transmission risk conditions to a savings of \$81,000 for surgeons under high prevalence/transmission risk conditions.

Given the high costs estimated for many of the options and the uncertainty of the results, the authors concluded as follows: "Given the ethical, social, and public health implications, mandatory testing policies should not be implemented without greater certainty as to their cost-effectiveness."

SOURCE: Adapted from Tevfik F. Nas, *Cost-Benefit Analysis: Theory and Application* (Thousand Oaks, CA: Sage, 1996), pp. 191-192. Original study was K. A. Phillips, R. A. Lowe, J. G. Kahn, P. Lurie, A. L. Avins, and D. Ciccarone, "The Cost Effectiveness of HIV Testing of Physicians and Dentists in the United States," *Journal of the American Medical Association*, 1994, 271:851-858.

in the estimates possible from available information. Given the high, but uncertain, cost estimates, policymakers would be wise to move cautiously on this issue until better information could be developed.

Because of the insufficient use of ex ante analyses in the social program arena, many social programs are initiated or markedly modified without attention to the practicality of the action in cost-benefit or cost-effectiveness

terms. For example, if the application of a particular dental treatment that prevents cavities costs \$200 per child annually, and the treatment is estimated to reduce cavities by an average of one-half cavity per child per year, it is unlikely to gain acceptance, even if it works. After all, its cost is four or five times what dentists would charge on average for filling a single cavity. An efficiency analysis in such a case might easily dissuade decisionmakers from implementing the program.

Most commonly, efficiency analyses in the social program field take place after the completion of an impact evaluation, when the net impact of a program is known. The focus of such *ex post* cost-benefit and cost-effectiveness assessments may be on examining the efficiency of a program in either absolute or comparative terms, or both. In all cases, the analysis is undertaken to assess whether the costs of the intervention can be justified by the magnitude of the net effects.

In absolute terms, the idea is to judge whether the program is worth what it costs by comparing costs either to benefits or to outcomes in substantive terms. For example, a cost-benefit analysis may reveal that for each dollar spent to reduce shoplifting in a department store, two dollars are saved in terms of stolen goods, an outcome that clearly indicates that the shoplifting program would be economically beneficial.

In comparative terms, the issue is to determine the differential "payoff" of one program versus another—for example, comparing the reduction in arrest rates for drunken driving brought about by an educational program with that of a program that pays for taxis to take people home after they have imbibed too much. In *ex post* analyses, estimates of costs and outcomes are based on studies of the types

described in previous chapters on program monitoring and impact evaluations.

The Concepts of Cost-Benefit and Cost-Effectiveness Analyses

Obviously, many considerations besides economic efficiency are brought to bear in policy making, planning, and program implementation, but economic efficiency is almost always critical, given that resources are inevitably scarce. Cost-benefit and cost-effectiveness analyses have the virtue of encouraging evaluators to become knowledgeable about program costs; surprisingly, many evaluators pay little attention to costs and are unaware of the information sources they need to contact and the complexities of describing program costs. In contrast, program costs are very salient to many of the stakeholder groups important to a program's acceptance and modification; consequently, attention to costs by evaluation staff often increases cooperation and support from such groups.

A cost-benefit analysis requires estimates of the benefits of a program, both tangible and intangible, and estimates of the costs of undertaking the program, both direct and indirect. Once specified, the benefits and costs are translated into a common measure, usually a monetary unit.

Cost-benefit analysis requires the adoption of a particular economic perspective; in addition, certain assumptions must be made to translate program inputs and outputs into monetary figures. As we have noted, there is considerable controversy in the field regarding the "correct" procedures to use in converting inputs and outputs into monetary values. Clearly, the assumptions underlying the defini-

tions of the measures of costs and benefits strongly influence the resulting conclusions. Consequently, the analyst is required, at the very least, to state the basis for the assumptions that underlie the analysis.

Often, analysts do more than that. They may undertake several different analyses of the same program, varying the assumptions made. For example, later we will discuss the need to take into account inflation (or deflation) in valuing costs and benefits that occur at different periods of time. The analyst could undertake a single study and state that an annual inflation rate of 5% was assumed, or the analyst could provide findings based on rates of 1%, 5%, and 10%. Sensitivity analyses, which alter important assumptions and estimate the consequences on program results, are a central feature of well-conducted efficiency studies. Indeed, an important advantage of formal efficiency studies over impressionistically gathered information about costs in relation to outcomes is that the assumptions and procedures are open to review and checking.

Cost-benefit analysis is least controversial when applied to technical and industrial projects, where it is relatively easy to place a monetary value on benefits as well as costs. Examples include engineering projects designed to reduce the costs of electricity to consumers, highway construction to facilitate transportation of goods, and irrigation programs to increase crop yields. Estimating benefits in monetary terms is frequently more difficult in social programs, where only a portion of program inputs and outputs may easily be assigned a monetary value. For example, it is possible to translate future occupational gains from an educational project into monetary values without incurring too much controversy. The issues are more complex in such social interventions as fertility control programs or

health services projects because one must ultimately place a value on human life to fully monetize the program benefits (Jones-Lee, 1994; Mishan, 1988).

The underlying principle is that cost-benefit analysts attempt to value both inputs and outputs at what is referred to as their marginal social values. For many items, such as the cost of providing a certain medicine or the monetary benefit of outfitting new cars with engines that burn less gasoline, market prices perform this task quite well. The situation is most difficult when the good or service is not even traded.

In general, there is much more controversy about converting outcomes into monetary values than there is about inputs. Because of the controversial nature of valuing outcomes, in many cases, especially regarding human services, cost-effectiveness analysis is seen as a more appropriate technique than cost-benefit analysis. Cost-effectiveness analysis requires monetizing only the program's costs; its benefits are expressed in outcome units. For example, the cost-effectiveness of distributing free textbooks to rural primary school children could be expressed in terms of how much each 1,000 project dollars increased the average reading scores of the targeted children.

For cost-effectiveness analysis, then, efficiency is expressed in terms of the costs of achieving a given result. That is, the efficiency of a program in attaining its goals is assessed in relation to the monetary value of the resources or costs put into the program for a designated unit of outcome. For example, alternative educational interventions may be compared by measuring the costs for each of achieving a specific educational gain as measured by test scores.

An example of relating costs to gains in mathematics and reading effects among elementary school children is shown in Exhibit

EXHIBIT 11-C Cost-Effectiveness of Computer-Assisted Instruction

To assist decisionmakers in considering different approaches to improving the mathematics and reading performance of elementary school children, a cost-effectiveness study was undertaken of computer-assisted instruction (CAI) compared to three alternative interventions. The findings run counter to some conventional expectations. Although the CAI alternative did relatively well according to the cost-effectiveness criterion, it did not do as well as peer tutoring. It is somewhat surprising that a traditional and a labor-intensive approach (peer tutoring) appears to be far more cost-effective than an electronic intervention, a widely used CAI approach. Moreover, the low ranking for the option of increasing the instructional time in the classroom, the centerpiece of

many of the calls for educational reform, makes it a relatively poor choice for both reading and mathematics from a cost-effectiveness perspective (see table).

To estimate the cost-effectiveness of the various alternatives, the researchers first determined the magnitude, in standard deviation units (effect sizes), of the increases on mathematics and reading achievement test scores resulting from each approach. They then determined the cost of each instructional approach and computed the achievement score effect per \$100 spent per student for each approach. The results, averaging the mathematics and reading achievement findings, are presented in the table.

Average Cost-Effectiveness Ratios of Four Interventions for Two Subject Areas (average of mathematics and reading effect sizes for each \$100 cost per student per subject)

Intervention	Cost-Effectiveness Ratio
Cross-age tutoring	
Combined peer and adult program	.22
Peer component	.34
Adult component	.07
Computer-assisted instruction	.15
Reducing class size	
From 35 to 30	.11
From 30 to 25	.09
From 25 to 20	.08
From 35 to 20	.09
Increasing instructional time	.09

SOURCE: Adapted from H. M. Levin, G. V. Glass, and G. R. Meister, "Cost-Effectiveness of Computer-Assisted Instruction," *Evaluation Review*, 1987, 11(1):50-72.

11-C. The analysis found that counseling by other students provided more impact per \$100 than other approaches. Surprisingly, such peer counseling was more cost-effective than a high-tech, computer-assisted instruction program.

An ex ante cost-effectiveness analysis allows potential programs to be compared and ranked according to the magnitudes of their effects relative to their estimated costs. In ex post cost-effectiveness analyses, actual pro-

gram costs and impact—and, hence, inputs and outputs—replace, to a considerable extent, estimates and assumptions. Moreover, retrospective analyses can yield useful insights and experiences, or methodological procedures that can be applied to future programs. However, comparisons of outcomes in relation to costs require that the programs under consideration have the same types of outcomes. If programs produce different outcomes, such as reduction in number of days in bed in the case of a medical care program and increased reading competence in the instance of an educational program, then one is still left with the difficulty of valuing the two outcomes. That is, how much is an average reduction of two bed days “worth” compared with a mean increase of 10 points on a standard reading test?

The Uses of Efficiency Analyses

Efficiency analyses, at least *ex post* analyses, can be considered an extension of, rather than an alternative to, impact evaluation. Since the estimation of either monetized benefits or substantive effects depends on knowledge of a program's *net* impact, it is impossible to engage in cost-benefit or cost-effectiveness calculations for programs in which impacts are unknown and inestimable. It also is senseless to do so for ineffective programs—that is, when impact evaluations discover no significant net effects. It is equally foolish to undertake efficiency analyses of ongoing or completed programs unless there are reasonable estimates of program impact.

When applied to efficacious programs, efficiency analyses are useful to those who must make policy decisions regarding the support of one program over another, or who need to decide in absolute terms whether the outcomes of a program are worth its costs, or who are

required to review the utility of programs at different points in time. Moreover, efficiency analysis can be useful in determining the degree to which different levels or “strengths” of interventions produce different levels of benefits and can be used in a formative manner to help improve program performance (Yates, 1996).

METHODOLOGY OF COST-BENEFIT ANALYSIS

To carry out a cost-benefit analysis, one must first decide which perspective to take in calculating costs and benefits. What point of view should be the basis for specifying, measuring, and monetizing benefits and costs? In short, costs to and benefits for whom? Benefits and costs must be defined from a single perspective because mixing points of view results in confused specifications and overlapping or double counting. Of course, several cost-benefit analyses for a single program may be undertaken, each from a different perspective. Separate analyses based on different perspectives often provide information on how benefits compare to costs as they affect relevant stakeholders.

Accounting Perspectives

Earlier, we referred to the need to identify an accounting perspective in estimating costs and benefits. Generally, three accounting perspectives may be used for the analysis of social projects, those of (a) individual participants or targets; (b) program sponsors; and (c) communal aggregates, or the society involved.

The *individual-target* accounting perspective takes the point of view of the units that are the program targets, that is, the persons, groups, or organizations receiving the interven-

tion or services. Cost-benefit analyses using the individual-target perspective often produce higher benefit-to-cost results (net benefits) than those using other perspectives. In other words, if the sponsor or society bears the cost and subsidizes a successful intervention, then the individual program participant benefits the most. For example, an educational project may impose relatively few costs on participants. Indeed, the cost to targets may primarily be the time spent in participating in the project, since books and materials usually are furnished. Furthermore, if the time required is primarily in the afternoons and evenings, there may be no loss of income involved. The benefits to the participants, meanwhile, may include improvements in earnings as a result of increased education, greater job satisfaction, and increased occupational options, as well as transfer payments received while participating in the project.

The *program sponsor* accounting perspective takes the point of view of the funding source in valuing benefits and specifying cost factors. The funding source may be a private agency or foundation, a government agency, or a for-profit firm. From this perspective, the cost-benefit analysis most closely resembles what frequently is termed *private profitability analysis*. That is, analysis from this perspective is designed to reveal what the sponsor will pay to provide a program and what benefits (or “profits”) should accrue to the sponsor.

The program sponsor accounting perspective is most appropriate when the sponsor must make decisive choices between alternative programs in the face of a fixed budget, that is, when there is no possibility of generating any additional funds. Under these circumstances, if, for example, the program sponsor is a county government, it may favor a vocational education initiative that includes student stipends over

other programs because this type of program would reduce the costs of public assistance and similar subsidies (since some of the persons in the vocational education program would have been supported by income maintenance funds). Also, if the future incomes of the participants were to increase because of the training received, their direct and indirect tax payments would increase, and these also could be included in calculating benefits from a program sponsor perspective. The costs to the government sponsor include the costs of operation, administration, instruction, supplies, facilities, and any additional subsidies or transfers paid to the participants during the training. As another illustration, Exhibit 11-D shows a cost-benefit calculation involving the savings to the mental health system that result from providing specialized services to patients with co-occurring mental disorders and substance abuse problems.

The *communal* accounting perspective takes the point of view of the community or society as a whole, usually in terms of total income. It is therefore the most comprehensive perspective, but also usually the most complex and thus the most difficult to apply. Taking the point of view of society as a whole implies that special efforts are being made to account for secondary or indirect project effects—effects on groups not directly involved with the intervention. Moreover, in the current literature, communal cost-benefit analysis has been expanded to include equity considerations, that is, the distributional effects of programs among different subgroups. From a communal standpoint, for example, every dollar earned by a minority member who had been unemployed for six months or more may be seen as a “double benefit” and so entered into the analyses.

Exhibit 11-E illustrates the benefits that need to be taken into account from a commu-

EXHIBIT 11-D Costs and Savings to the Mental Health System of Providing Specialized Dual Diagnosis Programs

People with serious mental disorders and co-occurring substance disorders (*dual diagnosis*) are very difficult and costly to treat in usual mental health or substance abuse services. Providing them with specialized dual diagnosis treatment programs might improve the outcomes but would add to the cost of services. However, if those improved outcomes decreased the need for subsequent mental health services they might result in savings that would offset the costs of the specialized program. Viewed from the perspective of policymakers in the mental health system, therefore, a crucial question is whether the cost to the mental health system for specialized programs for this client population will be recovered in savings to the system through reduced need for subsequent services.

To address this question, a team of evaluation researchers randomly assigned 132 patients to three specialized dual diagnosis programs and assessed both the outcomes and the costs. The "control" program was based on a 12-step recovery model and was the "usual care" condition for dual diagnosis patients in this mental health system. It involved referral to community Alcoholics Anonymous or Narcotics Anonymous meetings and associated supportive services to help the client manage the recovery process. A more intensive program option used a behavioral skills model that relied on cognitive-behavioral treatment focusing on social and independent living skills and relapse prevention. A less intensive option featured case management in which reduced caseloads allowed clinicians to

provide individualized assistance in such areas as daily living, housing, legal problems, and the like.

The behavioral skills model produced the largest positive effects on measures of client functioning and symptoms but was also the most expensive program to deliver. To further explore the cost considerations, the evaluators examined service utilization and cost data for the clients in each of the three programs for four time periods: the six months before the dual diagnosis programs began (baseline), the six months after, the 12 months after, and the 18 months after.

Mental health service costs were divided into two categories: supportive services and intensive services. Supportive services included case management, outpatient visits, medication visits, day services, and other such routine services for mental health patients. Intensive services included the more costly treatments for serious episodes, for instance, inpatient services, skilled nursing care, residential treatment, and emergency visits.

The costs of supportive services were expected to show an increase for all of the specialized dual diagnosis programs, corresponding to the extra resources required to provide them. Any significant savings to the mental health system were expected to appear as a result of decreased use of expensive intensive services. Thus, the cost analysis focused on the amount by which the costs of supportive services increased from baseline in comparison to the amount by which the costs of intensive services decreased. The table shows the results for the

EXHIBIT 11-D Continued

change in service utilization costs between the six-month baseline period and the 18 months after the program began.

As expected, the cost of supportive services generally increased after the specialized programs were implemented, except for the case management program, which actually showed a reduction in total support cost from the baseline service period. The largest increase in support costs, on the other hand, was associated with the relatively intensive behavioral skills program.

Also, as hoped, the costs for intensive services were reduced from baseline for all of the specialized programs. The greater impacts of the behavioral skills program on client functioning and symptoms, however, did not translate into corresponding decreases in service utilization and associated cost savings. Indeed, the usual-care condition of the 12-step program produced the greatest decreases in subsequent costs for intensive services. However, while the case management program did not yield such large decreases, its lower support costs resulted

in a savings-to-costs ratio that was comparable to that of the 12-step program. Additional analyses showed that these programs also generally resulted in savings to the medical system, the criminal justice system, and the families of the clients.

In terms of costs and savings directly to the mental health system, therefore, both the 12-step and the case management programs produced considerably more savings than they cost. Indeed, the cost analysis estimated that for every \$1 invested in providing these programs there were about \$9 in savings that would accrue over the subsequent 18 months. Moreover, the case management program could actually be implemented with a net reduction in support service costs, thus requiring no additional investment. The behavioral skills program, on the other hand, produced a net loss to the mental health system. For every \$1 invested in it, there was only a \$0.53 savings to the mental health system.

Average per Client Change in Costs of Services Used From Baseline to 18 Months Later, in Dollars

	12-Step Program	Behavioral Skills	Case Management
Change in mental health supportive costs (a)	+728	+1,146	-370
Change in mental health intensive costs (b)	-6,589	-612	-3,291
Ratio of (b) to (a)	9.05	0.53	8.89

SOURCE: Adapted from Jeanette M. Jerrell and Teh-Wei Hu, "Estimating the Cost Impact of Three Dual Diagnosis Treatment Programs," *Evaluation Review*, 1996, 20(2):160-180.

EXHIBIT 11-E Costs to Benefits of Correctional Sentences

The control of crime by appropriate sentencing of convicted offenders must take into account not only the costs of implementing each of the three choices typically available to judges—prison, jail, or probation sentences—but also the benefits derived. Each correctional approach generates different types of “benefits” for society. The major ones are *incapacitation* through removing the offender from the community by incarceration in a prison or jail, *deterrence* by making visible the consequences of criminal behavior to discourage potential offenders, and *rehabilitation* by resocialization and redirection of criminals’ behavior. Since jail sentences are usually short, for instance, the incapacitation benefit is very small compared with the benefit from prison sentences, although, since no one likes being in jail, the deterrence benefit of jail is estimated to be about five-sixths that of prison.

Estimated Annual Social Costs and Benefits per Offender, in Dollars, for Different Correctional Sentences (average across all offenses)

	Incapacitation Benefit	Rehabilitation Benefit	Deterrence Benefit	Costs	Net Benefits
Prison	+6,732	10,356	+6,113	-10,435	-7,946
Jail	+774	-5,410	+5,094	-2,772	-2,315
Probation	0	-2,874	+5,725	-1,675	+1,176

SOURCE: Adapted from T. Gray, C. R. Larsen, P. Haynes, and K. W. Olson, “Using Cost-Benefit Analysis to Evaluate Correctional Sentences,” *Evaluation Review*, 1991, 15(4):471-481.

nal perspective. In this exhibit, Gray and associates (1991) report on an effort to integrate several studies to come out with a reasonable cost-to-benefit analysis of the efficiency of different correctional approaches. As shown in the table in Exhibit 11-E, benefits are of several different types. Although, as the article carefully notes, there are limitations to the preci-

Gray and associates attempted to estimate the monetary value of these different social benefits for each sentencing option (see table).

While, on average, probation sentences showed greater net benefits than jail which, in turn, showed a smaller negative benefit than prison, the relative weight given to each benefit varied according to the type and circumstances of the offense. For example, the costs of a burglary (adding loss to the victim with costs of the police investigation, arrest, and court costs) comes to about \$5,000, suggesting that perhaps long prison sentences are called for in the case of recidivist burglars to maximize the incapacitation benefit. In contrast, the cost of apprehending and trying persons for receiving stolen property is less than \$2,000, and a short jail sentence or even probation may be the most efficient response.

sion of the estimates, the results are important to judges and other criminal justice experts concerned with the effects of different types of sentences.

The components of a cost-benefit analysis conducted from a communal perspective include most of the costs and benefits that also appear in calculations made from the individ-

ual and program sponsor perspectives, but the items are in a sense valued and monetized differently. For example, communal costs for a project include opportunity costs in terms of alternative investments forgone by the community to fund the project in question. These are obviously not the same as opportunity costs incurred by an individual as a consequence of participating in the project. Communal costs also include outlays for facilities, equipment, and personnel, usually valued differently than they would be from the program sponsor perspective. Finally, these costs do not include transfer payments because they would also be entered as benefits to the community and the two entries would simply cancel each other out.

Obviously, the decision about which accounting perspective to use depends on the stakeholders who constitute the audience for the analysis, or who have sponsored it. In this sense, the selection of the accounting perspective is a political choice. An analyst employed by a private foundation interested solely in containing the costs of hospital care, for example, often will take a program sponsor accounting perspective. The analyst may neglect or be uninterested in whether the cost-containment program that has the highest net benefits from a sponsor accounting perspective might actually show a negative cost-to-benefit value when viewed from the standpoint of the individual. This could be the case if the individual accounting perspective included the costs involved in having family members stay home from work because the early discharge of patients required them to provide the bedside care ordinarily received in the hospital.

Generally, the communal accounting perspective is the most politically neutral. If analyses using this perspective are done properly, the information gained from an individual or a program sponsor perspective will be included

as data about the distribution of costs and benefits. Another approach is to undertake cost-benefit analyses from more than one accounting perspective. The important point, however, is that cost-benefit analyses, like other evaluation activities, have political features.

Exhibit 11-F shows some of the basic components of cost-benefit analyses for the different accounting perspectives (the program sponsor in this case is a government agency). The list is not to be taken as complete but as an illustration only. Specific items included in real analyses vary.

Exhibit 11-G provides a simplified, hypothetical example of cost-benefit calculations for a training program from the three accounting perspectives. Again, the monetary figures are gross oversimplifications; a real analysis would require far more complex treatment of the measurement issues involved. Note that the same components may enter into the calculation as benefits from one perspective and as costs from another and that the difference between benefits and costs, or net benefit, will vary, depending on the accounting perspective used.

In some cases, it may be necessary to undertake a number of analyses. For example, if a government group and a private foundation jointly sponsor a program, separate analyses may be required for each to judge the return on its investment. Also, the analyst might want to calculate the costs and benefits to different groups of targets, such as the direct and indirect targets of a program. For example, many communities offer tax advantages to industrial corporations if they build their plants there; the intent is to provide employment opportunities for residents. Costs-to-benefits comparisons could be calculated for the employer, the employees, and also the “average” resident of the community, whose taxes may rise to take up

EXHIBIT 11-F Components of Cost-Benefit Analyses From Different Perspectives

	<i>Individual (targets)</i>	<i>Program Sponsor (government)</i>	<i>Communal (communities in general)</i>
Benefits	Increase in net earnings (after taxes) Additional benefits received (e.g., direct transfers, fringe and noneconomic benefits)	Increase in tax revenues Decrease in expenses of public assistance and other subsidies Value of work done within the project (salary and fringes at market costs)	Increase in gross earnings (before taxes) Increase in other income (e.g., fringe benefits, excluding direct transfers) Decrease in expenses of alternative projects no longer applicable Value of work done within the project (salary and fringes at market costs)
Costs	Opportunity costs (net earnings forgone) Loss of direct subsidies no longer applicable (alternative social programs) Costs related to participation (e.g., fees, materials)	Taxes lost Project costs (e.g., capital, administrative, instructional, direct subsidies)	Opportunity costs (gross earnings forgone) Project costs (excluding direct subsidies or transfer payments)

the slack resulting from the tax break to the factory owners. Other refinements might be included as well. For example, we excluded direct subsidies from the communal perspective, both as a cost and as a benefit, because they probably would balance each other out; however, under certain conditions it may be that the actual economic benefit of the subsidies is less than the cost.

Measuring Costs and Benefits

The specification, measurement, and valuation of costs and benefits—procedures that are central to cost-benefit analysis—raise two distinct problems: first, identifying and measuring

all program costs and benefits, and second, expressing all costs and benefits in terms of a common denominator, that is, translating them into monetary values.

The problem of identifying and measuring costs and benefits is most acute for ex ante appraisals, where often there are only speculative estimates of costs and impact. However, data often are limited in ex post cost-benefit analyses as well. For many social interventions, the information from an evaluation (or even a series of evaluations) may in itself prove insufficient for a retrospective cost-benefit analysis to be carried out. Thus, evaluations often provide only some of the necessary information, and the analyst frequently must use additional sources or judgments.

EXHIBIT 11-G Hypothetical Example of Cost-Benefit Calculation From Different Accounting Perspectives

<i>Benefits/Costs</i>			
			\$100,000
(1) Earnings improvement of trainees (before taxes)			
(2) Earnings improvement of trainees (after taxes)			80,000
(3) Value of work done in training period			10,000
(4) Project costs for facility and personnel			50,000
(5) Project costs for equipment and supplies			5,000
(6) Trainee stipends (direct transfer payments)			12,000
(7) Earnings forgone by trainees (before taxes)			11,000
(8) Earnings forgone by trainees (after taxes)			9,000
(9) Taxes lost: (7) - (8)			2,000
	<i>Individual</i>	<i>Program Sponsor</i>	<i>Communal</i>
Benefits	(2) 80,000 (6) 12,000 <hr/> 92,000	(1) - (2) 20,000 (3) 10,000 <hr/> 30,000	(1) 100,000 (3) 10,000 <hr/> 110,000
Costs	(8) 9,000 <hr/> 9,000	(4) 50,000 (5) 5,000 (6) 12,000 (9) 2,000 <hr/> 69,000	(4) 50,000 (5) 5,000 (7) 11,000 <hr/> 66,000
B/C ratio	$\frac{92,000}{9,000} = 10.22$	$\frac{30,000}{69,000} = .44$	$\frac{110,000}{66,000} = 1.69$
Net benefit^a	83,000	-39,000	44,000

a. Note that net social benefit can be split into net benefit for trainees plus net benefit for the government; in this case, the latter is negative: 44,000 = 83,000 + (-39,000).

The second problem in many social programs is the difficulty of translating benefits and costs to monetary units. Social programs frequently do not produce results that can be valued accurately by means of market prices. For example, many would argue that the benefits of a fertility control project, a literacy campaign, or a program providing training in improved health practices cannot be monetized in ways acceptable to the various stakeholders. What value should be placed on the embarrassment of an adult who cannot read? In such

cases, cost-effectiveness analysis might be a reasonable alternative, because such analysis does not require that benefits be valued in terms of money, but only that they be quantified by outcome measures.

Monetizing Outcomes

Because of the advantages of expressing benefits in monetary terms, a number of approaches have been specified for monetizing

outcomes or benefits (Thompson, 1980). Five frequently used ones are as follows:

1. *Money measurements.* The least controversial approach is to estimate direct monetary benefits. For example, if keeping a health center open for two hours in the evening reduces targets' absence from work (and thus loss of wages) by an average of ten hours per year, then, from an individual perspective, the annual benefit can be calculated by multiplying the average wage by ten hours by the number of employed targets.

2. *Market valuation.* Another relatively non-controversial approach is to monetize gains or impacts by valuing them at market prices. If crime is reduced in a community by 50%, benefits can be estimated in terms of housing prices through adjustment of current values on the basis of prices in communities with lower crime rates and similar social profiles.

3. *Econometric estimation.* A more complicated approach is to estimate the presumed value of a gain or impact in market terms. For example, the increase in tax receipts from greater business revenue due to a reduced fear of crime could be determined by calculating relevant tax revenues of similar communities with lower crime rates, and then estimating the tax receipts that would ensue for the community in question. Such estimation may require complex analytical efforts and the participation of a highly trained economic analyst.

Econometric analysis, especially when performed with refined contemporary multivariate techniques, is a popular choice because it can account for the other influences on the variable in question (in the preceding example, taxes lost because of fear of crime). The ana-

lytical effort required to do quality econometric work is certainly complex, and the assumptions involved are sometimes troublesome. However, econometric analysis, like all good methodological procedures, requires making assumptions explicit and therefore enables others to evaluate the analytical basis of the claims made.

4. *Hypothetical questions.* A quite problematic approach is to estimate the value of intrinsically nonmonetary benefits by questioning targets directly. For instance, a program to prevent dental disease may decrease participants' cavities by an average of one at age 40; thus, one might conduct a survey on how much people think it is worth to have an additional intact tooth as opposed to a filled tooth. Such estimates presume that the monetary value obtained realistically expresses the worth of an intact tooth. Clearly, hypothetical valuations of this kind are open to considerable skepticism.

5. *Observing political choices.* The most tentative approach is to estimate benefits on the basis of political actions. If state legislatures are consistently willing to appropriate funds for high-risk infant medical programs at a rate of \$40,000 per child saved, this figure could be used as an estimate of the monetary benefits of such programs. But given that political choices are complex, shifting, and inconsistent, this approach is generally very risky.

In summary, all relevant components must be included if the results of a cost-benefit analysis are to be valid and reliable and reflect fully the economic effects of a project. When important benefits are disregarded because they cannot be measured or monetized, the project may appear less efficient than it is; if certain costs are omitted, the project will seem more effi-

cient. The results may be just as misleading if estimates of costs or benefits are either too conservative or too generous. As a means of dealing with the problem, analysts often will value everything that can reasonably be valued and then list the things that cannot be valued. They will then estimate the value that would have to be placed on the nonmonetary benefits for the project to be a "go."

Shadow Prices

Benefits and costs need to be defined and valued differently, depending on the accounting perspective used. For many programs, however, the outputs simply do not have market prices (e.g., a reduction in pollution or the work of a homemaker), yet their value must be estimated. The preferred procedure is to use *shadow prices*, also known as *accounting prices*, to reflect better than do actual market prices the real costs and benefits to society. Shadow prices are derived prices for goods and services that are supposed to reflect their true benefits and costs. Sometimes it is more realistic to use shadow prices even when actual prices are available. For example, suppose an experimental program is implemented that requires a director who is knowledgeable about every one of the building trades. For the single site, the sponsors may be fortunate to find a retired person who is very interested in the program and willing to work for, say, \$30,000 per year. But if the program was shown to be a success through an impact evaluation and a cost-benefit analysis was undertaken, it might be best to use a shadow price of, say, \$50,000 for the director's salary, because it is very unlikely that additional persons with the non-monetary interests of the first director could be found (Nas, 1996).

Opportunity Costs

The concept of *opportunity costs* reflects the fact that resources generally are limited. Consequently, individuals or organizations choose from existing alternatives the ways these resources are to be allocated, and these choices affect the activities and goals of the decisionmakers. The cost of each choice can be measured by the worth of the forgone options.

Although this concept is relatively simple, the actual estimation of opportunity costs often is complex. For example, a police department may decide to pay the tuition of police officers who want to go to graduate school in psychology or social work on the grounds that the additional schooling will improve the officers' job performance. To have the money for this program, the department might have to keep its police cars an extra two months each. The opportunity costs could in this case be estimated by calculating the additional repair costs for the department's automobiles that would be incurred if the cars were replaced later. Since in many cases opportunity costs can be estimated only by making assumptions about the consequences of alternative investments, they are one of the controversial areas in efficiency analyses.

Secondary Effects (Externalities)

Projects may have external or *spillover* effects—that is, side effects or unintended consequences that may be either beneficial or detrimental. Because such effects are not deliberate outcomes, they may be inappropriately omitted from cost-benefit calculations if special efforts are not made to include them. A secondary effect of a training program, for example, might be the spillover of the training to relatives,

neighbors, and friends of the participants. Among the more commonly discussed negative external effects of industrial or technical projects are pollution, noise, traffic, and destruction of plant and animal life.

For many social programs, two secondary effects are likely: displacement and vacuum effects. For example, an educational or training project may produce a group of newly trained persons who enter the labor market, compete with workers already employed, and displace them (i.e., force them out of their jobs). Project participants may also vacate jobs held previously, leaving a vacuum that other workers might fill.

Secondary effects, or externalities, may be difficult to identify and measure. Once found, however, they should be incorporated into the cost-benefit calculations.

Distributional Considerations

Traditionally, judgments of the effectiveness of social interventions are predicated on the notion that an effective intervention makes at least one person better off and nobody worse off. In economics, this yardstick is called the *Pareto criterion*. Cost-benefit analysis, however, does not use the Pareto criterion, but rather the *potential Pareto criterion*. Under this criterion, the gains must potentially compensate for the losses, with something left over. That is, it is presumed—although not necessarily tested—that if the program's impact is estimated, more targets will be better off than worse off, or, more accurately, that the "balance" between total gains and total losses will be positive. This criterion may be very difficult to satisfy in social programs, however, particularly those that rely on income transfers. Lowering the minimum wage for teenagers, for

instance, may increase their employment at the cost of reducing work opportunities for older adults.

Often the concern is not simply with winners versus losers but with movement toward equity within a target population. This is particularly true in the case of programs designed to improve the general quality of life of a group or community. The basic means of incorporating equity and distributional considerations in the cost-benefit analysis involves a system of weights whereby benefits are valued more if they produce the anticipated positive effects. Thus, if a lowered minimum wage for teenagers decreases the family incomes of the moderately disadvantaged, the dollars gained and lost could be weighted differently, depending on the degree of disadvantage to the families. Some accomplishments are worth more than others to the community, both for equity reasons and for the increase in human well-being, and should therefore be weighted more heavily.

The weights to be assigned can be determined by the appropriate decisionmakers, in which case value judgments will obviously have to be made. They may also be derived through certain economic principles and assumptions. In any case, it is clear that weights cannot be applied indiscriminately. Analysts will undoubtedly develop further refinements as they continue to deal with the issue of distributional effects.

An intermediate solution to considerations of equity in cost-benefit analyses is to first test to see whether the costs and benefits of a program meet the potential Pareto criterion. If so, calculations can be undertaken for separate subgroups in the population. Such disaggregation might be done for separate income groups, for instance, or for students with different levels of achievement. Such distributional issues

are especially important in analyses of issues like the effects of schooling where costs are in part borne by taxpayers who do not receive direct benefits. Publicly supported education yields benefits primarily to those who have children in school and, disproportionately, to those who are less well off and, hence, pay lower taxes.

Discounting

Another major element in the methodology of efficiency analyses concerns the treatment of time in valuing program costs and benefits. Intervention programs vary in duration, and successful ones in particular produce benefits that are derived in the future, sometimes long after the intervention has taken place. The effects of many programs are expected to persist through the participants' lifetimes. Consequently, evaluators often must extrapolate into the future to measure impact and ascertain benefits, especially when program benefits are gauged as projected income changes for participants. In particular, ex ante appraisals often extrapolate into the future in carrying out a complete analysis. Otherwise, the evaluation would be based only on the restricted period of time for which actual program performance data are available.

Consequently, costs and benefits occurring at different points in time must be brought into a common measure or made commensurable. In other words, the time patterns for costs and benefits of a program must be taken into account. The applicable technique is known as *discounting* and consists of reducing costs and benefits that are dispersed through time to a common monetary base or adjusting them to their present values. For example, costs are usually highest at the beginning of an interven-

tion, when many of the resources must be expended; they either taper off or cease when the intervention ends. Even when a cost is fixed or a benefit is constant, increments of expenditures made or benefits derived at different points in time cannot be considered equivalent. Instead of asking, "How much more will my investment be worth in the future?" standard economic practice is to ask, "How much less are benefits derived in the future worth compared to those received in the present?" The same goes for costs. The answer depends on what we assume to be the rate of interest, or the discount rate, and the time frame chosen. Exhibit 11-H provides an example of discounting.

The choice of time period on which to base the analysis depends on the nature of the program and whether the analysis is ex ante or ex post. All else being equal, a program will appear more beneficial the longer the time horizon chosen.

There is no authoritative approach for fixing the discount rate. One choice is to fix the rate on the basis of the opportunity costs of capital, that is, the rate of return that could be earned if the funds were invested elsewhere. But there are considerable differences in opportunity costs depending on whether the funds are invested in the private sector, as an individual might do, or in the public sector, as a quasi-government body may decide it must. The length of time involved and the degree of risk associated with the investment are additional considerations.

The results of a cost-benefit analysis are thus particularly sensitive to the choice of discount rate. In practice, evaluators usually resolve this complex and controversial issue by carrying out discounting calculations based on several different rates. Furthermore, instead of

EXHIBIT 11-H Discounting Costs and Benefits to Their Present Values

Discounting is based on the simple notion that it is preferable to have a given amount of capital in the present rather than in the future. All else equal, present capital can be saved in a bank to accumulate interest or can be used for some alternative investment. Hence, it will be worth more than its face value in the future. Put differently, a fixed amount payable in the future is worth less than the same amount payable in the present.

Conceptually, discounting is the reverse of compound interest, since it tells us how much we would have to put aside today to yield a fixed amount in the future. Algebraically, discounting is the reciprocal of compound interest and is carried out by means of the simple formula

$$\text{Present value of an amount} = \frac{\text{Amount}}{(1 + r)^t}$$

Year				
1	2	3	4	5
\$1,000	\$1,000	\$1,000	\$1,000	\$1,000
$(1 + .10)^1$	$(1 + .10)^2$	$(1 + .10)^3$	$(1 + .10)^4$	$(1 + .10)^5$
= \$909.09	= \$826.45	= \$751.32	= \$683.01	= \$620.92

where r is the discount rate (e.g., .05) and t is the number of years. The total stream of benefits (and costs) of a program expressed in present values is obtained by adding up the discounted values for each year in the period chosen for study. An example of such a computation follows.

A training program is known to produce increases of \$1,000 per year in earnings for each participant. The earnings improvements are discounted to their present values at a 10% discount rate for five years.

Over the five years, total discounted benefits equal \$909.09 + \$826.45 + . . . + \$620.92, or \$3,790.79. Thus, increases of \$1,000 per year for the next five years are not currently worth \$5,000 but only \$3,790.79. At a 5% discount rate, the total present value would be \$4,329.48. In general, all else being equal, benefits calculated using low discount rates will appear greater than those calculated with high rates.

applying what may seem to be an arbitrary discount rate or rates, the evaluator may calculate the program's *internal rate of return*, or the value that the discount rate would have to be for program benefits to equal program costs.

A related technique, *inflation adjustment*, is used when changes over time in asset prices should be taken into account in cost-benefit calculations. For example, the prices of houses

and equipment may change considerably because of the increased or decreased value of the dollar at different times.

Ethical Issues in Setting Values

It is clear that with the many considerations involved there can be considerable disagreement on the monetary values to be placed

on benefits. The disputes that arise in setting these values underlie much of the conflict over whether cost-benefit analysis is a legitimate way of estimating the efficiency of programs. An interesting discussion of this matter is the article by Skaburskis (1987) in which he discusses the decision-making process in planning the BART transit system for the San Francisco Bay Area. As one illustration, in discussing the monetary values to be placed on the indirect effects of the new transportation system, he asks, "Is reduced air pollution worth 5, 10, or 15 cents to the average Bay Area resident?" (p. 605). It is the answer to this question that he says determines whether certain areas of the community are redeveloped.

Comparing Costs to Benefits

The final step in cost-benefit analysis consists of comparing total costs to total benefits. How this comparison is made depends to some extent on the purpose of the analysis and the conventions in the particular program sector. The most direct comparison can be made simply by subtracting costs from benefits. For example, a program may have costs of \$185,000 and its benefits are calculated at \$300,000; in this case, the net benefit (or profit, to use the business analogy) is \$115,000. Although generally more problematic, sometimes the ratio of benefits to costs is used rather than the net benefit. This measure is generally regarded as more difficult to interpret and should be avoided (Mishan, 1988).

In discussing the comparison of benefits to costs, we have noted the similarity to decision making in business. The analogy is real. In particular, in deciding which programs to support, some large private foundations actually

phrase their decisions in investment terms. They may want to balance a high-risk venture (i.e., one that might show a high rate of return but has a low probability of success) with a low-risk program (one that probably has a much lower rate of return but a much higher probability of success). Thus, foundations, community organizations, or government bodies might wish to spread their "investment risks" by developing a portfolio of projects with different likelihoods and prospective amounts of benefit.

Sometimes, of course, the costs of a program are greater than its benefits. In Exhibit 11-I, a cost-to-benefit analysis is presented that documents the negative results of a federal initiative to control noise. In this analysis, the costs of regulatory efforts to control the noise from motorcycles, trucks, and buses were estimated to be considerably higher than the benefits of the program. In the exhibit's table, the findings for truck and bus regulations are reported; note the negative values when benefits are subtracted from costs and the less than 1.0 values resulting when benefits are divided by costs. Of course, one can quarrel over the measure of benefits, which was simply the increase in property values resulting from a decline in decibels (dBAs) of noise. Nevertheless, according to Broder (1988), the analysis was a major reason why the Reagan administration abandoned the program.

It bears noting that sometimes programs that yield negative values are nevertheless important and should be continued. For example, there is a communal responsibility to provide for severely retarded persons, and it is unlikely that any procedure designed to do so will have a positive value (subtracting costs from benefits). In such cases, one may still want to compare the costs to benefits of different programs,

EXHIBIT 11-I A Study of the Birth and Death of a Regulatory Agenda

It has long been the case that, once funded, government programs are almost impossible to eliminate. Most organizations build up constituencies over the years that can be called on to protect them if threatened. Thus, it was particularly remarkable that the federal Office of Noise Abatement and Control (ONAC) at the Environmental Protection Agency (EPA) was disbanded during the Reagan administration, thus terminating a major social regulatory program without a public outcry.

Although the halt in the spread of inefficient noise regulation is one of few examples of lasting

relief from social regulation provided by the Reagan administration, a further irony is that much of the economic analysis that was at least partly instrumental was produced by the prior administration. Specifically, President Carter's Council of Economic Advisors and the Council on Wage and Price Stability, an agency disbanded by the Reagan administration, had produced several economic analyses for the public docket that were highly critical of the regulations, although it was the Reagan administration that acted on these analyses.

Cost-Benefit Analysis of Truck and Bus Noise Regulations

	Truck Noise Regulations		Bus Noise Regulations	
	83 dBAs	80 dBAs	83 dBAs	80 dBAs
Benefits (a)	1,056	1,571	66.2	188.5
Costs (b)	1,241	3,945	358.8	967.3
Net benefits (a) - (b)	-185	-2,374	-292.6	-778.8
Benefit-cost ratio (a)/(b)	.85	.40	.18	.19

NOTE: dBAs = decibels. Costs and benefits are in millions of 1978 dollars except for ratios.

SOURCE: Adapted from I. E. Broder, "A Study of the Birth and Death of a Regulatory Agenda: The Case of the EPA Noise Program," *Evaluation Review*, 1988, 12(3):291-309.

such as institutional care compared with home care.

When to Do Ex Post Cost-Benefit Analysis

Earlier in this chapter, we discussed the importance of undertaking ex ante analyses in developing programs that result in irrevocable or almost irrevocable commitments. We also indicated that many more ex ante analyses are

called for in the social program arena than are currently performed. Too often it is only after programs are put into place that policymakers and sponsors realize that the programs' costs compared to their benefits make them impractical to implement on a permanent basis.

In the case of ex post evaluations, it is important to consider a number of factors in determining whether to undertake a cost-benefit analysis. In some evaluation contexts, the technique is feasible, useful, and a logical com-

EXHIBIT 11-J Cotton Dust Regulation: An OSHA Success Story

In the late 1970s, the Occupational Safety and Health Administration (OSHA) took a major step in attempting to promote the health of workers in the textile industry, tightening its standard on cotton dust levels in textile plants. Because the OSHA cotton dust standard was widely believed to be ineffective, it became the target of a major political debate and a fundamental U.S. Supreme Court decision. However, the evidence indicates that the standard has had a significant beneficial effect on worker health, and at a cost much lower

than originally anticipated. For instance, data on the relationship between exposure to cotton dust and disease incidence, as well as the disability data and the evidence based on worker turnover, suggest that the risks of byssinosis (lung disease) have been reduced dramatically. The cost of eliminating even cases classified as "totally disabled" is less than \$1,500, and thus there is a strong economic basis for the enforcement of OSHA standards.

Estimated Reduction in Byssinosis Cases Associated With the Introduction of the Cotton Dust Standard

Type of Case	No. of Cases Reduced per Year, 1978-1982	Total No. of Cases Reduced per Year If Full Compliance
Byssinosis, Grades 1/2 and 1	3,517	5,047
Byssinosis over Grade 1	1,634	2,349
Partial disabilities	843	1,210
Total disabilities	339	487

SOURCE: Adapted, with permission, from W. K. Viscusi, "Cotton Dust Regulation: An OSHA Success Story?" *Journal of Policy Analysis and Management*, 1985, 4(3):325-343. Copyright © 1985, John Wiley & Sons, Inc.

ponent of a comprehensive evaluation; in others, its application may rest on dubious assumptions and be of limited utility.

Optimal prerequisites of an ex post cost-benefit analysis of a program include the following:

- The program has independent or separable funding. This means that its costs can be separated from other activities.
- The program is beyond the development stage, and it is certain that net effects are significant.
- Program impact and magnitude of impact are known or can be validly estimated.

- Benefits can be translated into monetary terms.

- Decisionmakers are considering alternative programs, rather than simply whether or not to continue the existing project.

Ex post efficiency estimation—both cost-benefit and cost-effectiveness analyses—should be components of many impact evaluations. In Exhibit 11-J, the impact of a program to replace machinery in cotton mills that causes an inordinate amount of dust is reported. Viscusi (1985) provides two sets of figures in the exhibit's table, showing the number of cases of byssinosis (lung disease) and of

long-term disabilities that were reduced by the initiative as well as the estimated number of cases that might have been reduced given full compliance with the program. His cost data indicate that even total disabilities are prevented for less than \$1,500, clearly an amount that the most conservative factory owner must acknowledge represents a saving compared to the spiraling costs of disability insurance of industrial plants. Merely presenting the information on the number of cases of lung disease that would be reduced by enforcing OSHA's standards—without demonstrating the comparatively low costs of the program—probably would not have had much impact on plant owners.

COST-EFFECTIVENESS ANALYSIS

Cost-benefit analysis allows evaluators to compare the economic efficiency of program alternatives, even when the interventions are not aimed at common goals. After initial attempts in the early 1970s to use cost-benefit analysis in social fields, however, some evaluators became uneasy about directly comparing cost-benefit calculations for, say, family planning to those for health, housing, or educational programs. As we have noted, sometimes it is simply not possible to obtain agreement on critical values—for example, on the monetary value of a life prevented by a fertility control project, or of a life saved by a health campaign—and then compare the results.

In contrast to cost-benefit analysis, cost-effectiveness analysis does not require that benefits and costs be reduced to a common denominator. Instead, the effectiveness of a program in reaching given substantive goals is related to

the monetary value of the costs. In cost-effectiveness analyses, programs with similar goals are evaluated and their costs compared. Thus, one can compare two or more programs aimed at lowering the fertility rate, or different educational methods for raising achievement levels, or various interventions to reduce infant mortality.

Cost-effectiveness analysis thus allows comparison and rank ordering of programs in terms of their costs for reaching given goals or the various inputs required for different degrees of goal achievement. But because the benefits are not converted to a common denominator, we cannot ascertain the worth or merit of a given intervention in monetary terms from such analyses. Likewise, we cannot determine which of two or more programs in different areas produces better returns. We can compare the relative efficiency of different programs only if they have the same or roughly similar goals and have the same outcome measures. In these analyses, efficiency is judged by comparing costs for units of outcome.

Cost-effectiveness analysis can be viewed as an extension of cost-benefit analysis to projects with multiple and noncommensurable goals. It is based on the same principles and uses the same methods as cost-benefit analysis. The assumptions of the method, as well as the procedures required for measuring costs and discounting, for example, are the same for both approaches. Therefore, the concepts and methodology introduced previously with regard to cost-benefit analysis can also be regarded as a basis for understanding the cost-effectiveness approach.

Cost-effectiveness analysis is a particularly good method for evaluating programs with similar outcomes without having to monetize the outcomes. Moreover, if a service or program

is known to produce positive outcomes, or presumed to, cost-effectiveness analysis may be conducted in terms of costs per client served. Identifying such *unit costs* makes it possible to compare the efficiency of different programs that provide similar services or different service components within a multiservice program. Exhibit 11-K provides an example of a cost analysis of this sort for methadone treatment programs for intravenous drug abusers. Of particular interest to the evaluators was the relative magnitude of the costs per client for an add-on employment training component compared with the costs of the standard program. However, the analysis was also able to reveal differences in costs per client across programs at four separate sites.

Although some sponsors and program staff are prejudiced against efficiency analyses because they deal chiefly with "dollars" and not "people," the approach that underlies them is no different from that of any stakeholder who needs to assess the utility of implementing or maintaining a program. Our world of limited resources, though often decried, nevertheless requires setting one program against another and deciding resource allocation. Competent efficiency analysis can provide valuable information about a program's economic potential or actual payoff and thus is important for program planning, implementation, and policy decisions, as well as for gaining and maintaining the support of stakeholders.

EXHIBIT 11-K Cost Analysis of Training and Employment Services in Methadone Treatment

Prior evaluation research has shown that vocational and employment counseling for drug users not only has positive effects on employment but also on drug use and criminality. Despite these encouraging signs, many drug treatment programs have reduced or eliminated vocational services due to changes in program emphasis or financial pressures. Against this background, a team of evaluators at Research Triangle Institute conducted cost analysis on four methadone maintenance programs with employment services components to help decision-makers explore the feasibility of a renewed emphasis on vocational services in substance abuse treatment.

The standard treatment in these programs involved methadone maintenance for intravenous drug users for as long as 12 months or more, random urine tests approximately once a month, monthly individual counseling sessions, and one to four group counseling sessions per month.

The Training and Employment Program component (TEP) of these programs included vocational needs assessment, location of existing training and employment programs suitable to the needs of methadone clients, and placement into training and jobs. Each program had an

on-site vocational specialist to work with both the drug counselors and the clients to identify and address vocational issues, provide job-related services, and maintain weekly contact with each assigned client.

Findings from a randomized impact assessment of the standard methadone treatment (STD) plus TEP compared with STD only showed that the methadone clients had high rates of unemployment and lacked vocational services and that TEP helped them access such services, obtain training, and reduce their short-term unemployment.

Given these positive findings, the critical practical question is how much the TEP component added to the cost of the standard treatment program. To assess this, the evaluators examined the total costs and cost per client of TEP in comparison to the analogous costs of the standard program without TEP for each of the four program sites. The main results are summarized in the table.

The results of this analysis indicated that the cost per client of the TEP component ranged from \$1,648 to \$2,215, amounts corresponding to between 42% and 50% of the cost of the standard methadone treatment without TEP.

EXHIBIT 11-K Continued

Annual Total and per Client Costs of Adding Training and Employment Program (TEP) Services Compared With the Costs of Standard (STD) Services

	Program A	Program B	Program C	Program D
Personnel	\$38,402	\$41,681	\$49,762	\$50,981
Support and supplies for vocational specialists	11,969	14,467	17,053	6,443
Travel	1,211	3,035	2,625	1,870
Other overhead	7,736	14,033	2,619	2,728
Total annual TEP cost	59,318	73,217	72,060	62,022
TEP clients served	36	38	43	28
Cost per client served	\$1,648	\$1,927	\$1,676	\$2,215
Total annual STD cost	\$819,202	\$1,552,816	\$2,031,698	\$1,531,067
STD clients served	210	400	573	300
STD cost per client	\$3,901	\$3,882	\$3,546	\$5,104
Total TEP cost/total STD cost	7.2%	4.7%	3.5%	4.1%
TEP per client/STD per client	42.2%	49.6%	47.3%	43.4%

Because many methadone maintenance clients are not appropriate for training and employment services, however, a TEP component will not be applicable to the entire caseload of the standard treatment program. When the incremental costs of adding a TEP component to the total program were figured, therefore, the results showed that

the TEP component added only 3.5% to 7.2% to the total program budget. In addition, the analysis showed different degrees of efficiency across programs in providing both TEP and standard services, as indicated in the varying costs per client.

SOURCE: Adapted from M. T. French, C. J. Bradley, B. Calingaert, M. L. Dennis, and G. T. Karantzios, "Cost Analysis of Training and Employment Services in Methadone Treatment," *Evaluation and Program Planning*, 1994, 17(2):107-120.

SUMMARY

- ☒ Efficiency analyses provide a framework for relating program costs to outcomes. Whereas cost-benefit analyses directly compare benefits to costs in commensurable (monetary) terms, cost-effectiveness analyses relate costs expressed in monetary terms to units of substantive results achieved.
- ☒ Efficiency analyses can be useful at all stages of a program, from planning through implementation and modification. Currently, ex post analyses are more commonplace than ex ante analyses in the social program arena because reasonably sound estimates of costs and benefits prior to program implementation are often lacking. Nevertheless, ex ante analyses should be undertaken more often than they are, particularly for programs that are expensive either to implement or to evaluate. Different sets of assumptions can create a range of analyses; one thing these analyses may reveal is the improbability of achieving the desired net benefits under any sensible set of assumptions.
- ☒ Efficiency analyses use different assumptions and may produce correspondingly different results depending on which accounting perspective is taken: that of individual targets or participants, program sponsors, or the community or society. Which perspective should be taken depends on the intended consumers of the analysis and thus involves political choice.
- ☒ Cost-benefit analysis requires that program costs and benefits be known, quantified, and transformed to a common measurement unit; that they be projected into the future to reflect the lifetime of a program; and that future benefits and costs be discounted to reflect their present values.
- ☒ Options for monetizing outcomes or benefits include money measurements, market valuation, econometric estimation, hypothetical questions asked of participants, and observation of political choices. Shadow, or accounting, prices are used for costs and benefits when market prices are unavailable or, in some circumstances, as substitutes for market prices that may be unrealistic.
- ☒ In estimating costs, the concept of opportunity costs allows for a truer estimate but can be complex and controversial in application.
- ☒ The true outcomes of projects include spillover and distributional effects, both of which should be taken into account in full cost-benefit analyses.

- ☒ Cost-effectiveness analysis is a feasible alternative to cost-benefit analysis when benefits cannot be calibrated in monetary units. It permits programs with similar goals to be compared in terms of their relative efficiency and can also be used to analyze the relative efficiency of variations of a program.
- ☒ Efficiency analyses can require considerable technical sophistication and the use of consultants. As a way of thinking about program results, however, they direct attention to costs as well as benefits and have great value for the evaluation field.

Primary dissemination	Dissemination of the detailed findings of an evaluation to sponsors and technical audiences.
Secondary dissemination	Dissemination of summarized, often simplified findings to audiences composed of stakeholders.
Policy significance	The significance of an evaluation's findings for policy and program development (as opposed to their statistical significance).
Policy space	The set of policy alternatives that are within the bounds of acceptability to policymakers at a given point in time.
Direct utilization	Explicit utilization of specific ideas and findings of an evaluation by decisionmakers and other stakeholders.
Conceptual utilization	Long-term, indirect utilization of the ideas and findings of an evaluation.

THE SOCIAL CONTEXT OF EVALUATION

In the preceding chapters, we have been concerned mainly with the technical aspects of conducting systematic evaluations. From the outset, however, we have asserted our view that evaluations involve more than simply using appropriate research procedures. Evaluation research is a purposeful activity, undertaken to affect policy development, to shape the design and implementation of social interventions, and to improve the management of social programs. In the broadest sense of politics, evaluation is a political activity.

There are, of course, intrinsic rewards for evaluators, who may derive great pleasure from satisfying themselves that they have done as good a technical job as possible—like artists whose paintings hang in their attics and never see the light of day, and poets whose penciled foolscap is hidden from sight in their desk drawers. But that is not really what it is all about. Evaluations are a real-world activity. In the end, what counts is the critical acclaim with which an evaluation is judged by peers in the field and the extent to which it leads to modified policies, programs, and practices—ones that, in the short or long term, improve the conditions of human life.

In this last chapter, we examine the current status of the field of evaluation research, with emphasis on the social context of evaluations. Certainly, compared with the late 1970s, when the first edition of *Evaluation* was published, there is considerably greater sophistication today among evaluators, not only on technical matters but also on the place of evaluation research in the policy and social program arena. (For an overview of the growth and change in the field, see Chelimsky and Shadish, 1997; Haveman, 1987; Shadish, Cook, and Leviton, 1991. For a different view of change in evaluation, see Guba and Lincoln, 1989.)

At the same time, strains and tensions persist about methodological matters, the education of evaluators, and organizational arrangements for the conduct of evaluations. Moreover, there are political and ideological issues concerning the social responsibility of evaluators that continue to confront the field, disagreement on the most effective ways to disseminate findings, and differences of opinion about the best strategies for maximizing the utility of evaluations.

We acknowledge, furthermore, that each evaluation has its unique features, requiring specially tailored solutions to the problems

encountered. The individuality of each evaluation makes it difficult to offer many "principles" about the conduct of evaluations. Nevertheless, the field is now mature enough that it is possible to offer reasonably sound observations about the state of the evaluation art, as well as general guidelines and advice on the conduct of the work. This chapter is based on an admixture of our own experiences and the writings of colleagues who have addressed the various interpersonal, political, and structural issues that surround doing evaluations.

With the experience of the past several decades, evaluators have become more humble about the potency of their efforts and have come to realize that social policy cannot be based on evaluation alone. Even the strongest proponents of the evaluation enterprise realistically acknowledge that its potential contributions to social policy are constrained by the range of competencies and self-interests of both the persons who undertake evaluations and the consumers of them, by diversity in styles of work and organizational arrangements, and by the political considerations and economic constraints that accompany all efforts at planned social change. Most important of all, in a democratic society, social change cannot be determined by the rule of experts but, rather, should be the outcome of processes that take into account the views of the various interests concerned.

In addition, evaluators, most of whom are convinced that social programs might improve the human condition, have been disappointed by finding out that many do not produce marked improvements and some are not effective. We have learned that designing effective programs and properly implementing them is very difficult. To many, it has not been an uplifting experience to have been the bearer of bad news.

Accordingly, evaluators have experienced the frustrations, feelings of inadequacy, and lack of self-esteem of all groups whose efforts often fall short of their hopes and aspirations. And their response has been the same as well: a great amount of introspection, a concerted effort to shift the blame to others, and an outpouring of verbal and written commentaries about the dismal state of social and human affairs, in particular the futility of developing and implementing effective and efficient interventions. Some social commentators have even blamed the reported failures of evaluation on the inability of current evaluation practices to recognize successful programs as such (Schorr, 1997).

It is evident that simply undertaking well-designed and carefully conducted evaluations of social programs by itself will not eradicate our human and social problems. But the contributions of the evaluation enterprise in moving us in the desired direction should be recognized. There is considerable evidence that the findings of evaluations do often influence policies, program planning and implementation, and the ways social programs are administered, sometimes in the short term and other times in the long term.

THE PURPOSEFULNESS OF EVALUATION ACTIVITIES

As we will discuss in a later section, evaluation practitioners are diverse in their disciplinary outlooks, their ideological and political orientations, and their economic and career aspirations. Despite this diversity, however, nearly all evaluators share a common perspective about the purposefulness of their work. The major

rationale for doing applied work is to have an impact on the actions and thinking of the broad classes of persons who affect social change, and who in their policy and action roles use the findings and conclusions provided by evaluators.

Evaluation activities logically fall under the general rubric of "applied" social research. Although the boundaries separating "basic" or "academic" research from applied research are not always perfectly clear, there are qualitative differences between them (Freeman and Rossi, 1984). Some of these we have discussed or alluded to earlier, as when we noted that evaluations need to be conducted so that they are "good enough" to answer the questions under study. This pragmatic standard, of course, contrasts with that used by basic researchers, who typically strive for the "best" methodology that can be used in carrying out their research. Of course, basic research is also constrained by resources so that compromises are often necessary.

Three additional distinctions between applied and basic research are important to understand. First, basic research typically is initiated to satisfy the intellectual curiosity of the investigator and to contribute to the knowledge base of a substantive area of interest to the researcher and his or her peers. Basic research is often directed to topics that are of central concern to the discipline in question. In contrast, applied work is undertaken because it might contribute to solving a practical problem. In the evaluation field, most often the impetus for undertaking work comes not from the evaluators themselves but from persons and groups who are concerned with a particular social problem. Thus, it is imperative that the evaluator understands the *social ecology* of the evaluation field. This is the first major topic that we take up in this chapter.

Second, basic researchers generally are trained in a single disciplinary orientation to which they remain committed throughout their careers. They typically draw on a narrow band of methodological procedures, and from one study to the next address a limited substantive domain. For example, an economist may make the costs of health care her area of expertise and consistently apply econometric modeling procedures to her chosen area of study. Similarly, a sociologist might primarily use participant observation as her method of choice and devote most of her career to the study of the educational professions.

In contrast, evaluators sometimes move from one program area to another, confronting diverse questions that typically require familiarity with a range of research methods and of a variety of substantive areas. For example, one of the authors has conducted evaluations of programs concerned with nutrition, crime prevention, effects of natural disasters, child abuse and neglect, normative consensus, and various levels of education, using methods that range from randomized experiments to large-scale cross-sectional studies and the statistical analysis of archived administrative records. Some evaluators specialize in one or a few program areas, combining in a very productive way their detailed substantive knowledge with their evaluation expertise. The fact that evaluators can often be confronted with widely different subject areas raises a number of issues about the training, outlook, and theoretical perspectives of evaluators in contrast to basic researchers and, more generally, about the profession of evaluation (Shadish and Reichardt, 1987). The evaluation profession is the second major topic in this chapter.

Third, although ethical concerns are important in both basic and applied research, they loom larger and are of greater societal impor-

tance in applied work. If a basic researcher violates professional standards, his discipline may suffer, but if an applied researcher crosses the line the effects might be felt by programs, the target populations involved, and the society as a whole. Accordingly, the third major topic of this chapter will be concerned with important ethical issues encountered in applied research.

Fourth, there is a major difference in the audiences for basic and applied work, and in the criteria for assessing its utilization. Basic researchers are most concerned with their peers' responses to their studies; utilization is judged by the acceptance of their papers in prestigious journals and the extent to which the research stimulates work by others. Applied researchers judge themselves, and are judged by the sponsors of their studies, on how much of a contribution they make to the development and implementation of policies and programs and, ultimately, to the resolution of social problems. Utilization of evaluation results, and ways to maximize it, constitute our final topic.

THE SOCIAL ECOLOGY OF EVALUATIONS

The likelihood of evaluations being used depends on evaluators' recognition that the key determinants of their utilization are the social and political contexts in which the evaluations are undertaken. Consequently, to conduct successful evaluations, evaluators need to continually assess the social ecology of the arena in which they work.

Sometimes the impetus and support for an evaluation come from the highest decision-making levels: Congress or a federal agency may mandate evaluations of innovative pro-

grams, as the Department of Health and Human Services did in the case of waivers given to states for innovative reforms in income maintenance programs (Gueron and Pauly, 1991), or the president of a large foundation may insist that the foundation's major social action programs be evaluated, as in the case of the supported housing programs of the Robert Wood Johnson Foundation (Rog et al., 1995). At other times, evaluation activities are initiated in response to requests from managers and supervisors of various operating agencies and focus on administrative matters specific to those agencies and stakeholders (Oman and Chitwood, 1984). At still other times, evaluations are undertaken in response to the concerns of individuals and groups in the community who have a stake in a particular social problem and the planned or current efforts to deal with it.

Whatever the impetus may be, evaluators' work is conducted in a real-world setting of multiple and often conflicting interests. In this connection, two essential features of the context of evaluation must be recognized: the existence of multiple stakeholders and the related fact that evaluation is usually part of a political process.

The Range of Stakeholders

In undertaking their studies, evaluators usually find a diversity of individuals and groups with interest in their work and its outcomes. These stakeholders may hold competing and sometimes combative views on the appropriateness of the evaluation work and whose interest will be affected by the outcome. To conduct their work effectively and contribute to the resolution of the issues at hand, evaluators must understand their relationships

with the stakeholders involved as well as the relationships between stakeholders.

The starting point for achieving this understanding is to recognize the range of stakeholders who directly or indirectly can affect the usefulness of evaluation efforts, both as evaluators go about doing their work and afterward in their responses to the product. This faces the lone evaluator situated in a single school, hospital, or social agency as well as those associated with evaluation groups in large organized research centers, federal and state agencies, or elite and community foundations.

In an abstract sense, every citizen who should be concerned with the efficacy and efficiency of efforts to improve social conditions has a stake in the outcome of an evaluation. In practice, of course, the stakeholder groups concerned with any given evaluation effort are more narrowly based, consisting of those who perceive direct and visible interests in the program. Within stakeholder groups, various stakeholders typically have different perspectives on the meaning and importance of an evaluation's findings. These disparate viewpoints are a source of potential conflict not only between stakeholders themselves but also between these persons and the evaluator. No matter how an evaluation comes out, there are some to whom the findings are good news and some to whom they are bad news.

To evaluate is to make judgments; to conduct an evaluation is to provide empirical evidence that can be used to substantiate judgments. The distinction between making judgments and providing information on which judgments can be based is useful and clear in the abstract, but often difficult to delineate in practice. No matter how well an evaluator's conclusions about the effectiveness of a program are grounded in rigorous research design and sensitively analyzed data, some

stakeholders are likely to perceive the results of an evaluation to be arbitrary or capricious judgments and to react accordingly.

Very little is known about how evaluation audiences are formed and activated. Nor is it completely clear how the interests of stakeholder groups are engaged and acted on by a given evaluation outcome. Perhaps the only reliable prediction is that the parties most likely to be attentive to an evaluation both during its conduct and after a report has been issued are the evaluation sponsors and program managers and staff. Of course, these are the groups who usually have the most at stake in the continuation of the program and whose activities are most clearly judged by the evaluation report.

The reactions of beneficiaries or targets of a program are especially problematic. In many cases, beneficiaries may have the strongest stake in an evaluation's outcome, yet they are often the least prepared to make their voices heard. Target beneficiaries tend to be unorganized and scattered in space; often they are poorly educated and unskilled in political communication. Sometimes they are reluctant even to identify themselves. When target beneficiaries do make themselves heard in the course of an evaluation, it is often through organizations who aspire to be their representatives. For example, homeless persons rarely make themselves heard in the discussion of programs directed at relieving their distressing conditions. But the National Coalition for the Homeless, an organization mainly composed of persons who themselves are not homeless, will often act as the spokesperson in policy discussions dealing with homelessness.

Consequences of Multiple Stakeholders

There are two important consequences of the phenomenon of multiple stakeholders.

First, evaluators must accept that their efforts are but one input into the complex mosaic from which decisions and actions eventuate. Second, strains invariably result from the conflicts in the interests of these stakeholders. In part, these strains can be eliminated or minimized by anticipating and planning for them; in part, they come with the turf and must be dealt with on an ad hoc basis or simply accepted and lived with.

The multiplicity of stakeholders for evaluations generates strains for evaluators in three main ways. First, evaluators are often unsure whose perspective they should take in designing an evaluation. Is the proper perspective that of the society as a whole, the government agency involved, the program staff, the clients, or one or more of the other stakeholder groups? For some evaluators, especially those who aspire to provide help and advice on fine-tuning programs, the primary audience often appears to be the program staff. For those evaluators whose projects have been mandated by a legislative body, the primary audience may appear to be the community, the state, or the nation as a whole.

It is important that the issue of which perspective to take in an evaluation is not understood as an issue of whose bias to accept. Perspective issues are involved in defining the goals of a program and deciding which stakeholder's concerns should be attended to. In contrast, bias in an evaluation usually means distorting an evaluation's design to favor findings that are in accord with some stakeholder's desires. Every evaluation is undertaken from some set of perspectives, but an ethical evaluator tries to avoid biasing evaluation findings in the design or analysis.

Some schools of evaluation strongly emphasize that certain perspectives should dominate in the conduct of evaluations. The "utili-

zation-focused evaluation" approach (e.g., Patton, 1997) asserts that evaluations ought to be designed to reflect the interests of "primary users," specifying methods for determining in specific cases who they may be. The advocates of "empowerment evaluation" (e.g., Fetterman, Kaftarian, and Wandersman, 1996) claim that the aim of evaluations should be to empower marginalized groups, usually the poor and minorities, adopting their perspectives and calling for the participation of such groups in the design and analysis of evaluations. It must be emphasized that neither of these two approaches is biased, in the sense used above.

Our own views on the perspectives from which evaluations are to be conducted are more agnostic. In the Chapter 11 discussion of the different accounting perspectives for conducting efficiency analyses, we noted that there is no one proper perspective but, rather, that different perspectives may be equally legitimate. The clients' or targets' perspective cannot claim any more legitimacy than that of the program or the government agency that funds the program. The responsibility of the evaluator is not to take one of the many perspectives as *the* legitimate one, but rather to be clear from which perspectives a particular evaluation is being undertaken while explicitly giving recognition to the existence of other perspectives. In reporting the results of an evaluation, an evaluator should state, for example, that the evaluation was conducted from the viewpoint of the program administrators while acknowledging that there also exist the alternative perspectives of the society as a whole and of the client targets.

In some evaluations, it may be possible to provide several perspectives on a program. For example, from the viewpoint of a target client, an income maintenance program may be judged as falling short of providing enough

dollars to satisfy basic needs, whereas from the perspective of a state legislature, the main purpose of the program is to facilitate the movement of clients off program rolls, a perspective that might view the low level of payment as a desirable incentive. From the viewpoint of income maintenance clients, a successful program may be one that provides payment levels sufficient to meet basic consumption needs of beneficiaries, whereas legislators may view a generous income maintenance program as fostering welfare dependency.

Second, the evaluator must realize that sponsors of evaluations may turn on evaluators when the results do not support the worth of the policies and programs they advocate. Although evaluators often anticipate negative reactions from other stakeholder groups, frequently they are unprepared for the responses of the sponsors of evaluations to findings that are contrary to what was expected or desired. Evaluators are in a very difficult position when this occurs. Losing the support of the evaluation sponsors, for example, may leave them open to attacks by other stakeholders, attacks they expected would be fended off by the sponsors. There are legitimate grounds for concern: Sponsors are a major source of referrals for additional work in the case of outside evaluators, and the providers of paychecks for inside ones. An illustration of the problem is provided in Exhibit 12-A, in which the findings of a study of the homeless of Chicago were severely challenged by advocacy stakeholders. (For a very different view of the same events, see Hoch, 1990.) The reactions of stakeholders in Chicago should not be taken as universal—there are many instances in which unwelcome findings are accepted and even acted on.

Third, misunderstandings may arise because of difficulties in communicating with different stakeholders. The vocabulary of the

evaluation field is no more complicated and esoteric than the vocabularies of the social sciences from which it is derived. But this does not make the vocabulary of evaluation understandable and accessible to lay audiences. To take a concrete illustration, the concept of *random* plays an important role in impact assessment. Technically, the concept has a precise, nonpejorative meaning, as shown in Chapter 7. In lay language, however, random often has connotations of *haphazard*, *careless*, *aimless*, *casual*, and so on—all of which have pejorative connotations. To advocate the random allocation of targets to experimental and control groups means something quite precise and delimited to evaluation researchers but may connote something very different to lay audiences. Thus, evaluators use the term *random* at their peril if they do not at the same time carefully specify its meaning.

It may be too much to expect an evaluator to master the subtleties of communication relevant to all the widely diverse audiences for evaluations. Yet the problem of communication remains an important obstacle to the understanding of evaluation procedures and the utilization of evaluation results. Evaluators are therefore well advised to anticipate the communication barriers in relating to stakeholders, a topic we will discuss more fully later in this chapter.

Disseminating Evaluation Results

For evaluation results to be used, they must be disseminated to and understood by major stakeholders and the general public. For our purposes, *dissemination* refers to the set of activities through which knowledge about evaluation findings is made available to the range of relevant audiences.

EXHIBIT 12-A The Consequences of Contrary Results

In the middle 1980s, the Robert Wood Johnson Foundation and the Pew Memorial Trust provided a grant to the Social and Demographic Institute at the University of Massachusetts to develop practical methods of undertaking credible enumerations of the homeless. The two foundations had just launched a program funding medical clinics for homeless persons, and an accurate count of the homeless was needed to assess how well the clinics were covering their clients.

Our findings concerning how many homeless were in Chicago quickly became the center of a controversy. The interests of the Chicago homeless were defended and advanced by the Chicago Coalition for the Homeless and by the Mayor's Committee on the Homeless, both composed of persons professionally and ideologically devoted to these ends. These two groups were consistently called on by the media and by public officials to make assessments of the status of the Chicago homeless. Their views about homelessness in essence defined the conventional wisdom and knowledge on this topic. In particular, a widely quoted estimate that between 20,000 and 25,000 persons were homeless in Chicago came from statements made by the Coalition and the Mayor's Committee.

At the outset, the Chicago Coalition for the Homeless maintained a neutral position toward our study. The study, its purposes, and its funding sources were explained to the coalition, and we asked for their cooperation, especially in connection with obtaining consent from shelter operators to interview their clients. The coalition neither endorsed our study nor condemned it, expressing some skepticism concerning our approach and especially about the operational definition of homelessness, arguing for a broader definition of homelessness that would en-

compass persons in precarious housing situations, persons living double-upped with families, single-room-occupancy renters, and so on.

When the data from Phase I were processed, we were shocked by the findings. The estimate of the size of the homeless population was many magnitudes smaller than the numbers used by the coalition: 2,344, compared to 20,000-25,000. Because we had anticipated a much larger homeless population, our sample of streets was too small to achieve much precision for such small numbers. We began to question whether we had made some egregious error in sample design or execution. Adding to our sense of self-doubt, the two foundations that had supported most of the project also began to have doubts, their queries fueled in part by direct complaints from the advocates for the homeless. To add to our troubles, the Phase I survey had consumed all the funds that our sponsors had provided, which were originally intended to support three surveys spread over a year. After checking over our Phase I findings, we were convinced that they were derived correctly but that they would be more convincing to outsiders if the study were replicated. We managed to convince our funding sponsors to provide more funds for a second survey that was designed with a larger sample of Chicago blocks than Phase I. The street sample was also supplemented by special purposive samples in places known to contain large numbers of homeless persons (bus, elevated, and subway stations; hospital waiting rooms; etc.) to test whether our dead-of-the-night survey time missed significant numbers of homeless persons who were on the streets during the early evening hours but had found sleeping accommodations by the time our interviewing teams searched sample blocks.

EXHIBIT 12-A Continued

When the data were in from Phase II, our calculated estimates of the average size of the nightly homeless in Chicago was 2,020 with a standard error of 275. Phase II certainly had increased the precision of our estimates but had not resulted in substantially different ones. Using data from our interviews, we also attempted to estimate the numbers of homeless persons we may have missed because they were temporarily housed, in jail, in a hospital, or in prison. In addition, we estimated the number of homeless children accompanying parents (we found no homeless children in our street searches). Adding these additional numbers of homeless persons to the average number who were nightly homeless as estimated from our Phase I and Phase II surveys, we arrived at a total of 2,722. This last estimate was still very far from the 20,000- to 25,000-person estimates of the Chicago Coalition.

Although the final report was distributed to the Chicago newspapers, television stations, and interested parties on the same date, somehow copies of the report had managed to get into the hands of the Coalition. Both major Chicago newspapers ran stories on the report, followed the next day by denunciatory comments from members of the Coalition. Despite our efforts to direct attention to the findings on the composition of the homeless, the newspapers headlined our numerical estimates. The comments from the coalition were harshly critical, claiming that our study was a serious disservice to the cause of the homeless and an attempt to lull public consciousness by severely (and deliberately) underestimating the number of homeless. Coalition comments included suggestions that the content

of the report was dictated by the Illinois Department of Public Aid, that the study was technically defective, and that our definition of the homeless omitted the thousands of persons forced to live with friends and relatives or in substandard housing conditions, or who negotiated sleeping arrangements every night.

Invited to give a presentation to the Mayor's Committee on the Homeless, I found my talk greeted by a torrent of criticism, ranging from the purely technical to the accusation of having sold out to the conservative forces of the Reagan administration and the Thompson Republican Illinois regime. But the major theme was that our report had seriously damaged the cause of homeless people in Chicago by providing state and local officials with an excuse to dismiss the problem as trivial. (In point of fact, the Illinois Department of Public Aid pledged to multiply its efforts to enroll homeless persons in the income maintenance programs the department administered.) Those two hours were the longest stretch of personal abuse I have suffered since basic training in the Army during World War II. It was particularly galling to have to defend our carefully and responsibly derived estimates against a set of estimates whose empirical footings were located in a filmy cloud of sheer speculation.

Almost overnight, I had become *persona non grata* in circles of homeless advocates. When I was invited by the Johnson Foundation to give a talk at a Los Angeles meeting of staff members from the medical clinics the foundation financed, no one present would talk to me except for a few outsiders. I became a nonperson wandering through the conference, literally shunned by all.

SOURCE: Adapted from Peter H. Rossi, "No Good Applied Research Goes Unpunished!" *Social Science and Modern Society*, 1987, 25(1):74-79.

Dissemination is a definite responsibility of evaluation researchers. An evaluation that is not made accessible to its audiences is clearly destined to be ignored. Accordingly, evaluators must take care in writing their reports and make provision for assuring that findings are delivered to major stakeholders.

Obviously, results must be communicated in ways that make them intelligible to the various stakeholder groups. External evaluation groups, in particular, generally provide sponsors with technical reports that include detailed and complete (not to mention honest) descriptions of the evaluation's design, data collection methods, analysis procedures, results, suggestions for further research, and recommendations regarding the program (in the case of monitoring or impact evaluations), as well as a discussion of the limitations of the data and analysis. Technical reports usually are read only by peers, rarely by the stakeholders who count. Many of these stakeholders simply are not accustomed to reading voluminous documents, do not have the time to do so, and might not be able to understand them.

For this reason, every evaluator must learn to be a "secondary disseminator." *Secondary dissemination* refers to the communication of research results and recommendations that emerge from evaluations in ways that meet the needs of stakeholders (as opposed to *primary dissemination*, which in most cases is the technical report). Secondary dissemination may take many different forms, including abbreviated versions of technical reports (often called executive summaries), special reports in more elaborate format that are issued regularly by either evaluation groups or the evaluation sponsors, memos, oral reports complete with slides, and sometimes even movies and videotapes.

The objective of secondary dissemination is simple: to provide results in ways that can be comprehended by the legendary "intelligent layperson," admittedly a figure sometimes as elusive as Bigfoot. Proper preparation of secondary dissemination documents is an art form unknown to most in the field, because few opportunities for learning are available during one's academic training. The important tactic in secondary communication is to find the appropriate style for presenting research findings, using language and form understandable to audiences who are intelligent but unschooled in the vocabulary and conventions of the field. *Language* implies a reasonable vocabulary level that is as free as possible from esoteric jargon; *form* means that secondary dissemination documents should be succinct and short enough not to be formidable. Useful advice for this process can be found in Torres, Preskill, and Piontek (1996). If the evaluator does not have the talents to disseminate his or her findings in ways that maximize utilization—and few of us do—an investment in expert help is justified. After all, as we have stressed, evaluations are undertaken as purposeful activities; they are useless unless used.

Evaluation as a Political Process

Throughout this book, we have stressed that evaluation results can be useful in the decision-making process at every point during a program's evolution and operations. In the earliest phases of program design, evaluations can provide basic data about social problems so that sensitive and appropriate services can be designed. While prototype programs are being tested, prospective evaluations may provide estimates of net effects to be expected when the program is fully implemented. After programs

have been in operation, evaluations can provide considerable knowledge about accountability issues. But this is not to say that what is useful in principle will automatically be understood, accepted, and used. At every stage, evaluation is only one ingredient in an inherently political process. And this is as it should be: Decisions with important social consequences should be determined in a democratic society by political processes.

In some cases, project sponsors may contract for an evaluation with the strong anticipation that it will critically influence the decision to continue, modify, or terminate a project. In those cases, the evaluator may be under pressure to produce information quickly, so that decisions can be made expeditiously. In short, evaluators may have a receptive audience. In other situations, evaluators may complete their assessments of an intervention only to discover that decisionmakers react slowly to their findings. Even more disconcerting are the occasions when a program is continued, modified, or terminated without regard to an evaluation's valuable and often expensively obtained information.

Although in such circumstances evaluators may feel that their labors have been in vain, they should remember that the decision-making process is indeed complex and that the results of an evaluation are only one of the elements in decision making. This point was clearly illustrated as long ago as 1915 in the controversy over the evaluation of the Gary plan in New York City, described in Exhibit 12-B.

The many parties involved in a human service program, including program sponsors, managers and operators, and sometimes the participants, often have very high stakes in the program's continuation, and their frequently unsupportable but enthusiastic claims may

count more heavily than the coolly objective results of an evaluation. Moreover, whereas the outcome of an evaluation is simply a single argument on one side or another, the outcome of typical American political processes may be viewed as a balancing of a variety of interests.

In any political system that is sensitive to weighing, assessing, and balancing the conflicting claims and interests of a number of constituencies, the evaluator's role is that of an expert witness, testifying to the degree of a program's effectiveness and bolstering that testimony with empirically based information. A jury of decisionmakers and other stakeholders may give such testimony more weight than uninformed opinion or shrewd guessing, but they, not the witness, are the ones who must reach a verdict. There are other considerations to be taken into account.

To imagine otherwise would be to see evaluators as having the power of veto in the political decision-making process, a power that would strip decisionmakers of their prerogatives. Under such circumstances, evaluators would become philosopher-kings whose pronouncements on particular programs would override those of all the other parties involved.

In short, the proper role of evaluation is to contribute the best possible knowledge on evaluation issues to the political process and not to attempt to supplant that process. Exhibit 12-C contains an excerpt from an article by one of the founders of modern evaluation theory, Donald T. Campbell, expounding a view of evaluators as servants of "the Experimenting Society."

Political Time and Evaluation Time

There are two additional strains involved in doing evaluations, compared with academic

EXHIBIT 12-B Politics and Evaluation

This exhibit concerns the introduction of a new plan of school organization into the New York City schools in the period around World War I. The so-called Gary plan modeled schools after the new mass production factories, with children being placed on shifts and moved in platoons from subject matter to subject matter. The following account is a description of how evaluation results entered into the political struggle between the new school board and the existing school system administration.

The Gary plan was introduced into the schools by a new school board appointed by a reform mayor, initially on a pilot basis. School Superintendent Maxwell, resentful of interference in his professional domain and suspicious of the intent of the mayor's administration, had already expressed his feelings about the Gary plan as it was operating in one of the pilot schools: "Well, I visited that school the other day, and the only thing I saw was a lot of children digging in a lot." Despite the superintendent's views, the Gary system had been extended to 12 schools in the Bronx, and there were plans to extend it further. The cry for more research before extending the plan was raised by a school board member.

social research, that are consequences of the fact that the evaluator is engaged in a political process involving multiple stakeholders: One is the need for evaluations to be relevant and significant in a policy sense, a topic we will take up momentarily; the other is the difference between political time and evaluation time.

Evaluations take time, especially those directed at assessing program impact. Usually, the tighter and more elegant the study design, the longer the time period required to perform

In the summer of 1915, Superintendent Maxwell ordered an evaluative study of the Gary plan as it had been implemented in the New York schools. The job was given to B. R. Buckingham, an educational psychologist in the research department of the New York City schools and a pioneer in the development of academic achievement tests. Buckingham used his newly developed academic achievement tests to compare two Gary-organized schools, six schools organized on a competing plan, and eight traditionally organized schools. The traditionally organized schools came out best on average, while the two Gary-organized schools averaged poorest.

Buckingham's report was highly critical of the eager proponents of the Gary system for making premature statements concerning its superiority. No sooner had the Buckingham report appeared than a veritable storm of rebuttal followed, both in the press and in professional journals. Howard W. Nudd, executive director of the Public Education Association, wrote a detailed critique of the Buckingham report, which was published in the *New York Globe*, the *New York Times*, *School and Society*, and the *Journal of Education*. Nudd argued that at the time Buckingham

the evaluation. Large-scale social experiments that gauge the effects of major innovative programs may require anywhere from four to eight years to complete and document. The political and program worlds often move at a much faster pace. Policymakers and project sponsors usually are impatient to know whether or not a program is achieving its goals, and often their time frame is a matter of months, not years.

For this reason, evaluators frequently encounter pressure to complete their assessments

EXHIBIT 12-B Continued

conducted his tests, the Gary plan had been in operation in one school for only four months and in the other for less than three weeks. He asserted that much of the requested equipment had not been provided and that the work of the Gary schools had been seriously disturbed by the constant stream of visitors who descended to examine the program. In a detailed, school-by-school comparison, Nudd showed that in one of the Gary-organized schools 90% of the pupils came from immigrant homes where Italian was their first tongue while some of the comparison schools were largely populated by middle-class, native-born children. Moreover, pupils in one of the Gary schools had excellent test scores that compared favorably with those from other schools. When scores were averaged with the second Gary school, however, the overall result put the Gary plan well behind.

Buckingham had no answer to the contention of inadequate controls, but he argued that he was

dealing, not with two schools, six schools, or eight schools, but with measurements on more than 11,000 children and therefore his study represented a substantial test of the Gary scheme. He justified undertaking his study early on the grounds that the Gary plan, already in operation in 12 Bronx schools, was being pushed on the New York schools and superintendent precipitously. As noted above, there was pressure from the mayor's office to extend the plan throughout the New York City schools and to make any increase in the education budget contingent on wholesale adoption of the Gary system. The president of the Board of Education found it advantageous to cite Nudd's interpretation of the Buckingham report in debate at the Board of Education meeting. Superintendent Maxwell continued to cite the Buckingham study as evidence against the effectiveness of the Gary plan, even a year and a half later.

SOURCE: Adapted from A. Levine and M. Levine, "The Social Context of Evaluation Research: A Case Study," *Evaluation Quarterly*, 1977, 1(4):515-542.

more quickly than the best methods permit, as well as to release preliminary results. At times, evaluators are asked for their "impressions" of effectiveness, even when they have stressed that such impressions are liable to be useless in the absence of firm results. For example, evaluators are now being asked by the mass media and legislators how effective the welfare reforms initiated by the Personal Responsibility and Work Opportunity Act of 1996 are, although for almost all states, the reforms have

yet to be worked out in detail, much less been put in place. At the time of this writing (1998), no evidence exists on this topic. So great is the desire for evaluative evidence that the mass media relies on anecdotes, dramatic specific examples, and even wild guesses.

In addition, the planning and procedures related to initiating evaluations within organizations that sponsor such work often make it difficult to undertake timely studies. In most cases, procedures must be approved at several

EXHIBIT 12-C Social Scientists as Servants of the Experimenting Society

Societies will continue to use preponderantly unscientific political processes to decide upon ameliorative program innovations. Whether it would be good to increase the role of social science in deciding on the content of the programs tried out is not at issue here. The emphasis is rather more on the passive role for the social scientist as an aid in helping society decide whether or not its innovations have achieved desired goals without damaging side effects. The job of the methodologist for the experimenting society is not to say *what is to be done*, but rather to say *what has been done*. The aspect of social science that is being applied is primarily its research methodology rather than its descriptive theory, with the goal of learning more than we do now from the innovations decided upon by the political process. . . . This emphasis seems to be quite different from the present role as government advisors of most economists,

international relations professors, foreign area experts, political scientists, sociologists of poverty and race relations, psychologists of child development and learning, etc. Government asks what to do, and scholars answer with an assurance quite out of keeping with the scientific status of their fields. In the process, the scholar-advisors too fall into the overadvocacy trap and fail to be interested in finding out what happens when their advice is followed. Certainty that one already knows precludes finding out how valid one's theories are. We social scientists could afford more of the modesty of the physical sciences, [and] should more often say that we can't know until we've tried. . . . Perhaps all I am advocating . . . is that social scientists avoid cloaking their recommendations in a specious pseudo-scientific certainty, and instead acknowledge their advice as consisting of but wise conjectures that need to be tested in implementation.

SOURCE: Quoted, with permission, from Donald T. Campbell, "Methods for the Experimenting Society," *Evaluation Practice*, 1991, 12(3):228-229.

levels and by a number of key stakeholders. As a result, it can take considerable time to commission and launch an evaluation, not counting the time it takes to implement and complete it. Although both governmental and private sector sponsors have tried to develop mechanisms to speed up the planning and procurement processes, these efforts are hindered by the workings of their bureaucracies, legal requirements related to contracting, and the need to establish agreement on the evaluation questions and design.

It is not clear what can be done to reduce the pressure resulting from the different time

schedules of evaluators and decisionmakers. Obviously, a long-term study should not be undertaken if the information is needed before the evaluation can be completed. It may be better in such circumstances to rely on expert opinion or another of the more judgmental evaluation methods discussed in Chapter 10. At times, it is a judgment call whether it is better to have some information or no information at all. At the very least, it is important that evaluators anticipate the time demands of stakeholders, particularly the sponsors of evaluations, and avoid making unrealistic time commitments.

A strategic approach is to confine technically complex evaluations to pilot or prototype projects for interventions that are unlikely to be implemented on a large scale in the immediate future. Thus, randomized controlled experiments may be most appropriate to evaluate the worth of new programs (initially implemented on a relatively small scale) before such programs appear on the agendas of decision-making bodies.

Another strategy for evaluators is to anticipate the direction of programs and policy activities, rather than be forced to undertake work that cannot be accomplished in the time allocated. One proposal that has attracted some attention is to establish independent evaluation institutes dedicated to examining, on a pilot or prototype basis, interventions that might one day be in demand. Evaluation centers could be established that continually assess the worth of alternative social programs for dealing with social problems that are of perpetual concern or that have a high probability of emerging in the years ahead. Although this proposal has some attractive features, especially to professional evaluators, it is not at all clear that it is possible to forecast accurately what, say, the next decade's social issues will be.

Perhaps the most successful approximation of efforts to maximize the contributions of evaluation activities prior to the implementation of new initiatives is the prospective evaluation synthesis of the Program Evaluation and Methodology Division of the General Accounting Office (GAO). As Chelimsky (1987) describes in Exhibit 12-D, her division's *ex ante* activities can make important contributions to shaping social legislation. (See also Chelimsky, 1991, for a general view of how applied social research intersects with policy making.)

As things stand now, however, we believe that the tension caused by the disparities between political and research time will continue to be a problem in the employment of evaluation as a useful tool for policymakers and project managers.

Issues of Policy Significance

Evaluations, we have stressed, are done with a purpose that is practical and political in nature. In addition to the issues we have already reviewed, the fact that evaluations are ultimately conducted to affect the *policy-making* process introduces several considerations that further distinguish an evaluator's work from that of a basic researcher.

Policy Relevance and Policy Space

Policy space is that set of alternative policies that can garner political support at any given point in time. The alternatives considered in designing, implementing, and assessing a social program are ordinarily those that are within current policy space. Policy space keeps changing in response to the efforts of influential figures to gain support from other influential and from ordinary community members. This decade's policy space with respect to crime control is dominated by programs of long and sometimes mandatory sentences for selected types of criminals. In contrast, during the 1970s it was centered on the development of community-based treatment centers as an alternative to imprisonment, on the grounds that prisons were breeding places for crime and that criminals would be best helped by being kept in close contact with the normal, civilian world.

The volatility of policy space is illustrated by the Transitional Aid to Released Prisoners

EXHIBIT 12-D Using Evaluative Activities in the Analysis of Proposed New Programs

Many of us spend much of our time doing retrospective studies; these are and will continue to be the meat and potatoes of evaluation research. Congress asks us for them and asks the executive branch to do them, and they are needed, but these studies are not the easiest ones to insert into the political process, and they may well be the least propitious from the viewpoint of use. . . . By contrast, before a program has started, evaluators can have an enormous effect in improving the reasoning behind program purposes or goals, in identifying the problems to be addressed, and in selecting the best point of intervention and the type of intervention most likely to succeed. The tempo at which new programs are sometimes introduced presents some difficulty. . . . The pace often becomes so frantic that the lead time necessary to gear up for evaluative work is simply impossible to obtain if results are to be ready soon enough to be useful.

At the GAO we are developing a method I call the Evaluation Planning Review which is specifically intended to be useful in the formulation of new programs. We have just given it a first try by

looking at a proposed program focusing on teenage pregnancy. Essentially, the method seeks to gather information on what is known about past, similar programs and apply the experience to the architecture of the new one. Senator Chaffee asked us to look at the bill he was introducing; we managed to secure four good months to do the work, and it has been a major success from both the legislative point of view and our own. From a more general, political perspective, providing understanding ahead of time of how a program might work can render a valuable public service—either by helping to shore up a poorly thought-out program or by validating the basic soundness of what is to be undertaken. True, there are questions that decisionmakers do not pose to evaluators that could usefully be posed, which seems a priori to be a problem for the framework; however, even when evaluators have been free to choose the questions, this particular type of question has not often been asked. Also, evaluators can always influence the next round of policy questions through their products.

SOURCE: Eleanor Chelimsky, "The Politics of Program Evaluation," *Society*, 1987, 25(1):26-27. Reprinted by permission of Transaction Publishers. Copyright 1987 by Transaction Publishers; all rights reserved.

(TARP) experiments, discussed in earlier chapters, which were conducted in the late 1970s to evaluate the effectiveness in reducing recidivism of providing short-term financial support to recently released felons. Whatever the merits of the Georgia and Texas TARP experiments, by the time the evaluation findings were available, federal policy space had changed so drastically that the policies emerging from those experiments had no chance of being consid-

ered. Thus, evaluators need to be sensitive not only to the policy space that exists when a research program is initiated but also to ongoing changes in the social and political context that alter the policy space as the evaluation proceeds.

Too often a prospective program may be tested without sufficient understanding of how the policy issues are seen by those decisionmakers who will have to approve the enact-

ment of the program. Hence, even though the evaluation of the program in question may be flawless, its findings may prove irrelevant. In the New Jersey-Pennsylvania income maintenance experiment, the experiment's designers posed as their central issue the following question: How large is the work disincentive effect of an income maintenance plan? By the time the experiment was completed and congressional committees were considering various income maintenance plans, however, the key issue was not the work disincentive effect—the policy space had changed. Rather, members of Congress were more concerned with how many different forms of welfare could be consolidated into one comprehensive package, without ignoring important needs of the poor and without creating many inequities (Rossi and Lyall, 1976).

Because a major purpose of impact assessments, as with evaluative activities generally, is to help decisionmakers form and adopt social policies, the research must be sensitive to the various policy issues involved and the limits of policy space. The goals of a project must resemble those articulated by policymakers in deliberations on the issues of concern. A carefully designed randomized experiment showing that a reduction in certain regressive taxes would lead to an improvement in worker productivity may be irrelevant if decisionmakers are more concerned with motivating entrepreneurs and attracting potential investments.

For these reasons, responsible impact assessment design must necessarily involve, if at all possible, some contact with relevant decisionmakers to ascertain their interests in the project being tested. A sensitive evaluator needs to know what current and future policy space will allow consideration. For an innovative project that is not currently being discussed by decisionmakers but is being tested because

it may become the subject of future discussion, the evaluators and sponsors of the test of impact effectiveness must rely on their informed guesses about what policy issues might arise, that is, what are likely prospective changes in policy space. For other projects, the processes of obtaining decisionmakers' opinions are quite straightforward. Evaluators can consult the proceedings of deliberative bodies (e.g., government committee hearings or legislative debates), interview decisionmakers' staffs, or consult decisionmakers directly.

Interpreting evaluation results involves considerations that go beyond methodology. The fact that evaluations are conducted according to the canons of social research may make them superior to other modes of judging social programs. But evaluations provide only superfluous information unless they address the value issues of persons engaged in policy making, program planning, and management.

The weaknesses of evaluations, in this regard, tend to center on how research questions are stated and how findings are interpreted (Datta, 1980). To maximize the utility of evaluation findings, evaluators must be sensitive to two levels of policy considerations.

First, programs that address problems perceived as critical require better (i.e., more rigorous) assessments than interventions related to relatively trivial concerns. Technical decisions, such as setting levels of statistical significance and magnitude, should be informed by the nature of policy and program considerations. These are always matters of judgment and sensitivity. Even when formal efficiency analyses (Chapter 11) are undertaken, the issue remains. For example, the decision to use an individual, program, or community accounting perspective is determined by policy and sponsorship considerations. Second, evaluation findings have to be assessed according to how

far they are generalizable, whether the findings are significant to the policy and to the program, and whether the program clearly fits the need (as expressed by the many factors that are involved in the policy-making process).

Policy Significance Versus Statistical Significance

An evaluation may produce results that all would agree are statistically significant and generalizable and yet be too small to be relevant to policy, planning, and managerial action (Lipsey, 1990; Sechrest and Yeaton, 1982). What the magnitude of a difference must be to have policy significance varies from field to field and from instance to instance. One formal way of providing data for such judgments is to conduct cost-benefit and cost-effectiveness analyses, as discussed in the previous chapter. Doing so allows judgments to be made on the basis of whether resources are effectively expended as compared to the costs and benefits of alternative projects, criteria that might not be appropriate for some programs. Other supplements to statistical inference tests have been proposed that involve taking into account the preponderance of evidence and replications (Browner and Newman, 1989; Goodman and Royall, 1988).

Another, more diffuse, criterion is to make judgments of the social worth of the change in outcome. Small magnitudes of change can have policy significance when the social worth of the change is high, correspondingly, large changes can have significance even when social worth is low. Thus, a program of nutritional education that reduces severe cases of malnutrition in children by only 2% undoubtedly would be regarded as policy-significant because malnutrition is regarded as a dangerous threat to children; a consumer education project that reduces the purchase of unnecessary small

household appliances by the same percentage probably would not be so regarded because consumer profligacy is not regarded as very serious. However, if the consumer education program reduced such purchases by 20%, it probably would be seen as policy-significant.

The availability of alternative interventions also needs to be taken into account. For example, in a country highly saturated with television sets and with a formal educational system that can be modified only over a long time period and through the expenditure of extensive resources, small gains from educational television may be significant for policy. The same magnitude of change would not be viewed positively if rapid changes at low cost were possible in the formal educational system.

Basic Science Models Versus Policy-Oriented Models

Social scientists often do not grasp the difference in emphasis required in formulating a model purposefully to *alter* a phenomenon as opposed to developing a causal model to *explain* the phenomenon. For example, much of the criminal behavior of young men can be explained by the extent of such behavior among males in their social network—fathers, brothers, other male relatives, friends, neighbors, schoolmates, and so on. This is a fascinating finding that affords many insights into the geographic and ethnic distributions of crime rates. However, it is not a useful finding in terms of altering the crime rate because it is difficult to envisage an acceptable public policy that would alter the social networks of young men. Short of yanking young males out of their settings and putting them into other environments, it is not at all clear that anything can be done to affect their social networks. Policy

space will likely, and, it is hoped, never include population redistribution for these purposes.

In contrast, although a weaker determinant of crime, it is easier to envisage a public policy that would attempt to alter the perceived costs of engaging in criminal activities. For example, altering potential lawbreakers' subjective probabilities of being caught for committing a crime, being convicted if caught, and going to prison if convicted can be a practical basis for a program of crime control. The willingness to engage in crime is sluggishly and weakly related to these subjective probabilities: The more that individuals believe they likely will be caught if they commit a crime, convicted if caught, and imprisoned if convicted, the lower the probability of criminal behavior. Thus, to some extent the incidence of criminal acts will be reduced if the police are effective in arresting criminals, if the prosecution is diligent in obtaining convictions, and if the courts have a harsh sentencing policy. None of these relationships is especially strong, yet these findings are much more appropriate to public policy that attempts to control crime than the social network explanation discussed earlier. Mayors and police chiefs can implement programs that increase the proportion of criminals apprehended, prosecutors can work harder at obtaining convictions, and judges can refuse to plea-bargain. Moreover, dissemination of these policy changes in ways that reach the potential offenders would, in itself, have some modest impact on the crime rate. The general point should be clear: Basic social science models often ignore policy-relevance.

The Missing Engineering Tradition

Our discussion of policy-relevant and policy-significant research points to a more general lesson: In the long term, evaluators—in-

deed, all applied researchers—and their stakeholders must develop an "engineering tradition," something that currently is missing in most of the social sciences. Engineers are distinguished from their "pure science" counterparts by their concern with working out the details of how scientific knowledge can be used to grapple with real-life problems. It is one thing to know that gases expand when heated and that each gas has its own expansion coefficient; it is quite another to be able to use that principle to mass-produce economical, high-quality gas turbine engines.

Similar engineering problems exist with respect to social science findings. It is well known in social science theories in economics and in psychological learning theory that changing incentives can often alter behavior. In the 1980s, there developed a fair amount of consensus that the incentives involved in welfare payments under Aid to Families With Dependent Children (AFDC) fostered dependency and hindered the movement of AFDC clients off the rolls into employment. Accordingly, the Department of Health and Human Services encouraged states to modify AFDC rules to provide incentives for clients to seek and obtain employment. Several versions of incentive packages were tested in randomized experiments. The experiments tested programs in which adults on welfare were prepared through training for employment, allowed to retain some proportion of their earnings without reduction in welfare payments, and aided to find employment. The findings of the experiments were that aiding in the employment search was more effective than training and that the combination of the two strategies was the most effective (Guéron and Pauly, 1991).

We are not certain how such social science engineers should be trained, and we suspect that training models will have to await the

appearance of a sufficient number of exemplars to learn from. Our hope is that the foregoing observations about the dynamics of conducting evaluations in the context of the real world of program and social policy sensitize the evaluator to the importance of "scouting" the terrain when embarking on an evaluation and of remaining alert to ecological changes that occur during the evaluation process. Such efforts may be at least as important to the successful conduct of evaluation activities as the appropriateness of the technical procedures employed.

Evaluating Evaluations

As evaluations have become more sophisticated, judging whether some particular evaluation was performed skillfully and findings interpreted properly becomes more and more difficult. Especially for laypersons and public officials, assessing the credibility of evaluations may be beyond their reach. In addition, there may often be contradictory research findings arising from several evaluations of the same program: How to reconcile conflicting evaluation claims can present problems even to evaluation experts. To meet the need for validating evaluations and for adequate communication of their findings, several approaches have been tried, as discussed below.

Quite frequently, the contracts or grants funding large-scale evaluations call for the formation of advisory committees composed of evaluation experts and policy analysts to oversee the conduct of the evaluation and provide expert advice to the evaluators and the funders. The advisory committee approach can be viewed as a way to raise the quality of evaluations and at the same time to provide greater legitimacy to their findings.

There also have been intensive reviews of evaluations, including reanalyses of evaluation datasets. For example, the National Academy of Sciences from time to time forms committees to review evaluations and synthesize their findings on topics of policy interest or significant controversy. For example, Coyle, Boruch, and Turner (1991) reviewed AIDS education evaluations with regard to their findings and also recommended improvements in the quality of such work.

Reviews such as those mentioned above typically take several years to complete and hence do not meet the needs of policymakers who require more timely information. More timely commentary on evaluations requires more rapid review and assessment. A promising attempt to be timely was funded in 1997 through a grant from the Smith-Richardson Foundation. The University of Maryland's School of Public Affairs was commissioned to convene a "blue ribbon" commission of prominent evaluators and policy analysts to review and comment on the expected considerable flow of evaluations of the reforms in public welfare undertaken under the Personal Responsibility and Work Opportunity Reconciliation Act of 1996. The Committee to Review Welfare Reform Research will issue periodic reports addressed to policymakers assessing the adequacy of the evaluations and drawing out their implications for policy. It is planned for the evaluation reviews to appear within a few months after the release of evaluation reports (Besharov, Germanis, and Rossi, 1998).

Despite these examples, we believe that the typical program evaluation is not ordinarily subject to the judgment of peers in the evaluation community. Some policymakers may have the competence to judge their adequacy, but most may have to rely on the persuasive qualities of evaluation reports. For this reason,

as discussed in a later section, evaluation standards are of recurring importance in the professional associations of evaluators.

THE PROFESSION OF EVALUATION

There is no single roster of all persons who identify themselves as evaluators and no way of fully describing their backgrounds or the range of activities in which they are engaged. At a minimum, some 50,000 persons are engaged, full- or part-time, in evaluation activities. We arrived at this estimate by adding together the numbers of federal, state, county, and city governmental organizations engaged in social program development and implementation, along with the numbers of school districts, hospitals, mental hospitals, and universities and colleges, all of which are usually obligated to undertake one or more types of evaluation activities. We do not know the actual number of persons engaged in evaluation work in these groups, and we have no way of estimating the numbers of university professors and persons affiliated with nonprofit and for-profit applied research firms who do evaluations. Indeed, the actual number of full- and part-time evaluators may be double or triple our minimum estimate. It is clear that evaluators work in widely disparate social program areas and devote varying amounts of their working time to evaluation activities. At best, the role definition of the evaluator is blurred and fuzzy.

At the one extreme, persons may perform evaluations as an adjunct activity. Sometimes they undertake their evaluation activities simply to conform to legislative or regulatory requirements, as apparently is the case in many local school systems. To comply with state or federal funding requirements, schools must

have someone designated as an "evaluator," and so name someone on their teaching or management staffs to serve in that capacity. Often the person appointed has no particular qualifications for the assignment either by training or by experience. Indeed, sometimes the appointee is someone who is not highly regarded as either a teacher or an administrator and is given evaluation duties as a harmless assignment.

At the other extreme, within university evaluation institutes and social science departments, and within applied social research firms in the private and nonprofit sectors, there are full-time evaluation specialists, highly trained and with years of experience, who are working at the frontiers of the evaluation field.

Indeed, the common labeling of persons as evaluators or evaluation researchers conceals the heterogeneity, diversity, and amorphousness of the field. Evaluators are not licensed or certified, so the identification of a person as an evaluator provides no assurance that he or she shares any core knowledge with another person so identified. The proportion of evaluators who interact and communicate with each other, particularly across social program areas, likely is very small. The American Evaluation Association, the major "general" organization in the field, has only a few thousand members, and the cross-disciplinary journal with the most subscribers, *Evaluation Review*, is read by only a few thousand. Within program areas, there likewise are only weak social networks of evaluators, most of whom are unaffiliated with national and local professional organizations that have organized evaluation "sections."

In brief, evaluation is not a "profession," at least in terms of the formal criteria that sociologists generally use to characterize such groups. Rather, it can best be described as a "near-group," a large aggregate of persons who

are not formally organized, whose membership changes rapidly, and who have little in common with one another in terms of the range of tasks they undertake or their competencies, work sites, and outlooks. This feature of the evaluation field underlies much of the discussion that follows.

Intellectual Diversity and Its Consequences

All the social science disciplines—economics, psychology, sociology, political science, and anthropology—have contributed to the development of the field of evaluation. Persons trained in each of these disciplines have made contributions to the conceptual base of evaluation research and to its methodological repertoire. The human service professional fields have also made contributions: Persons trained in the various human service professions with close ties to the social sciences—medicine, public health, social welfare, urban planning, public administration, education, and so on—have made important methodological contributions and have undertaken landmark evaluations. In addition, the applied mathematics fields of statistics, biometrics, econometrics, and psychometrics have contributed important ideas on measurement and analysis.

Cross-disciplinary borrowing has been extensive. Take the following examples: Although economics traditionally has not been an experimentally based social science, economists have designed and implemented a significant proportion of the large-scale, randomized field experiments of the past several decades, including the highly visible employment training, income maintenance, housing allowance, and national health insurance experiments. Sociologists and psychologists have borrowed heav-

ily from the econometricians, notably in their use of time-series analysis methods and simultaneous equation modeling. Sociologists have contributed many of the conceptual and data collection procedures used in monitoring organizational performance, and psychologists have contributed the idea of regression-discontinuity designs to time-series analyses. Psychometricians have provided some of the basic ideas underlying theories of measurement applicable to all fields, and anthropologists have provided some of the basic approaches used in qualitative fieldwork. Indeed, the vocabulary of evaluation is a mix from all of these disciplines. The list of references at the back of this book is testimony to the multidisciplinary character of the evaluation field.

In the abstract, the diverse roots of the field are one of its attractions. In practice, however, they confront evaluators with the need to be general social scientists and lifelong students if they are even to keep up, let alone broaden their knowledge base. Furthermore, the diversity in the field accounts to a considerable extent for the "improper" selection of research approaches for which evaluators are sometimes criticized. Clearly, it is impossible for every evaluator to be a scholar in all of the social sciences and to be an expert in every methodological procedure.

There is no ready solution to the need to have the broad knowledge base and range of competencies ideally required by the "universal" evaluator. This situation means that evaluators must at times forsake opportunities to undertake work because their knowledge bases may be too narrow, that they may have to use an "almost good enough" method rather than the appropriate one they are unfamiliar with, and that sponsors of evaluations and managers of evaluation staffs must be highly selective in deciding on contractors and in

making work assignments. It also means that at times evaluators will need to make heavy use of consultants and solicit advice from peers.

In a profession, a range of opportunities is provided for keeping up with the state of the art and expanding one's repertoire of competencies—for example, the peer learning that occurs at regional and national meetings and the didactic courses provided by these professional associations. At present, only a fraction of the many thousands of evaluation practitioners participate in professional evaluation organizations and can take advantage of the opportunities they provide.

There also are liabilities to becoming a profession. Established professions can suffer from over-professionalization, as we know from the state of many of the practicing professions. But the near-group character of the field and its diverse roots have their consequences as well, consequences that are exacerbated by the different ways in which evaluators are educated.

The Education of Evaluators

Few people in evaluation have achieved responsible posts and rewards by working their way up from lowly jobs within evaluation units. Most evaluators have some sort of formal graduate training either in social science departments or in professional schools. One of the important consequences of the multidisciplinary character of evaluation is that appropriate training for full participation in it cannot be adequately undertaken within any single discipline. In a few universities, interdisciplinary programs have been set up that include graduate instruction in a number of departments. In these programs, a graduate student might be directed to take courses in test construction and measurement in a department of psychol-

ogy, econometrics in a department of economics, survey design and analysis in a department of sociology, policy analysis in a political science department, and so on.

Interdisciplinary training programs, however, are neither common nor very stable. In the typical graduate-training and research-oriented university, the powerful units are the traditional departments. The interdepartmental coalitions of faculty that form interdisciplinary programs tend to have short half-lives, because departments typically do not reward participation in such ventures very highly and faculty drift back into their departments as a consequence. The result is that too often graduate training of evaluators primarily is unidisciplinary despite the clear need for it to be multidisciplinary.

Moreover, even within academic departments, applied work is often regarded less highly than "pure" or "basic" research. As a consequence, training in evaluation-related competencies is often limited. Psychology departments may provide fine courses on experimental design but fail to consider very much the special problems of implementing field experiments in comparison with laboratory studies; sociology departments may teach survey research courses but not deal at all with the special data collection problems involved in interviewing the unique populations that are typically the targets of social programs. Then, too, the low status accorded applied work in graduate departments often is a barrier to undertaking evaluations as dissertations and theses. If there is any advice to be given, it is that individual students who are interested in an evaluation career must be assertive. Often the student must take the lead in hand-tailoring an individual study program that includes course offerings in a range of departments, be insistent about undertaking an applied dissertation or

thesis, and seize on any opportunities within university research institutes and in the community to supplement formal instruction with relevant apprenticeship learning.

The other training route is the professional school. Schools of education train evaluators for positions in that field, programs in schools of public health and medical care produce persons who engage in health service evaluations, and so on. In fact, over time these professional schools, as well as MBA programs, have become the training sites for many evaluators.

These programs have their limitations as well. One criticism raised about them is that they are too "trade school" oriented in outlook. Consequently, some of them fail to provide the conceptual breadth and depth that allows graduates to move back and forth across social program areas, and to grasp technical innovations when they occur. Moreover, particularly at a master's level, many professional schools are required to have a number of mandatory courses, because their standing and sometimes their funding depend on accreditation by professional bodies who see the need for common training if graduates are going to leave as MSWs, MPHs, MBAs, and the like. Because many programs therefore leave little time for electives, the amount of technical training that can be taken in courses is limited. Increasingly, the training of evaluators in professional schools therefore has moved from the master's to the doctoral level.

Also, in many universities both faculty and students in professional schools are viewed as second-class citizens by those located in social science departments. This elitism often isolates students so that they cannot take advantage of course offerings in several social science departments or apprenticeship training in their affiliated social science research institutes. Students trained in professional schools, particu-

larly at the master's level, often trade off opportunities for intensive technical training for substantive knowledge in a particular program area and the benefits of professional certification. The obvious remedy is either undertaking further graduate work or seizing opportunities for additional learning of technical skills while pursuing an evaluation career.

We hold no brief for one route over the other; each has its advantages and liabilities. Increasingly, it appears that professional schools are becoming the major suppliers of evaluators, in part at least because of the reluctance of graduate social science departments to develop appropriate applied research programs. But these professional schools are far from homogeneous in what they teach, particularly in the methods of evaluation they emphasize, thus, the continued diversity of the field.

Consequences of Diversity in Origins

The existence of many educational pathways to becoming an evaluator contributes to the lack of coherence in the field. It accounts, at least in part, for the differences in the very definition of evaluation, and the different outlooks regarding the appropriate way to evaluate a particular social program. Of course, other factors contribute to this diversity, including social and political ideologies of evaluators.

Some of the differences are related to whether the evaluator is educated in a professional school or a social science department. For example, evaluators who come out of professional schools such as social work or education are much more likely than those trained in, say, sociology to see themselves as part of the program staff and to give priority to tasks that help program managers. Thus, they are likely to stress *formative* evaluations that are

designed to improve the day-to-day operations of programs.

The diversity is also related to differences among social science departments and among professional schools. Evaluators trained as political scientists frequently are oriented to *policy analysis*, an activity designed to aid legislators and high-level executives, particularly government administrators. Anthropologists, as one might expect, are predisposed to qualitative approaches and are unusually attentive to target populations' interests in evaluation outcomes. Consonant with their discipline's emphasis on small-scale experiments, psychologists often are concerned more with the validity of the causal inference in their evaluations than the generalizability to program practice. In contrast, sociologists are often more concerned with the potential for generalization and are more willing to forsake some degree of rigor in the causal conclusions to achieve it. Economists are likely to work in still different ways, depending on the body of microeconomic theory to guide their evaluation designs.

Similar diversity can be found among those educated in different professional schools. Evaluators trained in schools of education may focus on educational competency tests in measuring the outcome of early-childhood education programs, whereas social work graduates focus on caseworker ratings of children's emotional status and parental reports of their behavior. Persons coming from schools of public health may be most interested in preventive practices, those from medical care administration programs in frequency of physician encounters and duration of hospitalization, and so on.

It is easy to exaggerate the distinctive outlook that each discipline and profession manifests in approaching the design and conduct of

evaluations and there are many exceptions to the preference tendencies just described. Indeed, a favorite game among evaluation buffs is to guess an author's disciplinary background from the content of an article he or she has written. Nevertheless, disciplinary and professional diversity has produced a fair degree of conflict within the field of evaluation. Evaluators hold divided views on topics ranging from epistemology to the choice of methods and the major goals of evaluation. Some of the major divisions are described briefly below.

Orientations Toward Primary Stakeholders

As mentioned earlier in this chapter, some evaluators believe that evaluations should be mainly directed toward helping program managers to improve their programs. This view of evaluation sees its purpose primarily as consultation to program management to the point that the difference between technical assistance and evaluation becomes blurred. According to this view, an evaluation succeeds to the extent that programs are improved. These evaluation orientations tend also to avoid making judgments about the worth of programs on the grounds that most programs can be made to work with the help of evaluators. (See Patton, 1997, for a prominent advocate of utilization-focused evaluation.)

Others hold that the purpose of evaluations should be to help program beneficiaries (targets) to become empowered. This view of evaluation believes that engaging targets in a collaborative effort to define programs and their evaluation leads targets to become more "in charge" of their lives and leads to an increase in the sense of personal efficacy. (Fetterman, Kaftarian, and Wandersman, 1996, contains examples of this approach.)

At the other extreme of this division are the evaluators who believe that evaluators should mainly serve those stakeholders who fund the evaluation. Such evaluations take on the perspective of the funders, adopting their definitions of program goals and program outcomes.

Our own view has been stated earlier in this chapter. We believe that evaluations ought to be sensitive to the perspectives of major stakeholders. Ordinarily, the contractual requirements ruling evaluations require that primary attention be given to the evaluation sponsor's definitions of program goals and outcomes. However, such requirements do not exclude other perspectives. We believe that it is the obligation of evaluators to state clearly the perspective from which the evaluation is undertaken and to point out what other major perspectives are involved. When an evaluation has the resources to accommodate several perspectives, multiple perspectives should be used.

Epistemological Differences

The "cultural wars" that have been waged in the humanities and some of the social sciences have also touched evaluation as well. Postmodern theories of knowledge claim that positivistic epistemology has been superseded by a more relativistic view of knowledge and are reflected in evaluation with claims that social problems are social constructions and that knowledge is not absolute but that there are different "truths," each valid for the perspectives from which it derives. Postmodernists tend to favor qualitative research methods that produce rich "naturalistic" data and evaluation perspectives favoring those of the program personnel and target populations. (See Guba and Lincoln, 1989, for a foremost exponent of postmodern evaluation.)

The epistemological contrast to the postmodern position is not homogeneous in its beliefs on the nature of knowledge. Nevertheless, there is some strong consensus that truth is not entirely relativistic. For example, most believe that the definition of poverty is a social construction, but at the same time, there is the conviction that the distribution of annual incomes can be described through research operations on which most social scientists can agree. That is, whether a given income level is regarded as poverty is a matter of social judgment but the number of households at that income level can be estimated with a known sampling error. This position implies that disagreements among researchers on empirical findings are mainly matters of method or measurement error rather than matters involving different truths.

Our own position, as exemplified throughout this book, is clearly not postmodern. We believe that there are close matches between methods and evaluation problems. For given research questions, there are better methods and poorer methods. Indeed, the major message in this book is how to choose the best method for a given research question that is likely to produce the most credible findings.

The Qualitative-Quantitative Division

Coinciding with some of the divisions within the evaluation community is the division between those who advocate qualitative methods and those who argue for quantitative ones. A sometimes pointless literature has developed around this. On one side, the advocates of qualitative approaches stress the need for intimate knowledge and acquaintance with a program's concrete manifestations in attaining valid knowledge about the program's effects.

Qualitative evaluators tend to be oriented toward formative evaluation, that is, making a program work better by feeding information on the program to its managers. In contrast, quantitatively oriented evaluators often view the field as one primarily concerned with summative evaluation and focus on developing measures of program characteristics, processes, and impact that allow program effectiveness to be assessed with relatively high credibility.

Often the polemics obscure the critical point—namely, that each approach has utility, and the choice of approaches depends on the evaluation question at hand. We have tried in this volume to identify the appropriate applications of each viewpoint. As we have stressed, qualitative approaches can play critical roles in program design and are important means of monitoring programs. In contrast, quantitative approaches are much more appropriate in estimates of net impact as well as in assessments of the efficiency of social program efforts. (For a balanced discussion of the qualitative-quantitative discussion, see Reichardt and Rallis, 1994.)

Thus, it is fruitless to raise the issue of which is the better approach without specifying the evaluation questions to be studied. Fitting the approach to the research purposes is the critical issue: To pit one approach against the other in the abstract results in a pointless dichotomization of the field. Even the most avid proponents of one approach or the other recognize the contribution each makes to social program evaluations (Cronbach, 1982; Patton, 1997). Indeed, the use of multiple methods, often referred to as *triangulation*, can strengthen the validity of findings if results produced by different methods are congruent. Using multiple methods is a means of offsetting different kinds of bias and measurement error (for an

extended discussion of this point, see Greene and Caracelli, 1997).

The problem, as we see it, is both philosophical and strategic. Evaluations are undertaken primarily as contributions to policy and program formulation and modification—activities, as we have stressed, that have a strong political dimension. As Chelmsky (1987) has observed, "It is rarely prudent to enter a burning political debate armed only with a case study" (p. 27).

Diversity in Working Arrangements

The diversity of the evaluation field is also manifest in the variety of settings and bureaucratic structures in which evaluators work. First, there are two contradictory theses about working arrangements, or what might be called the insider-outsider debate. One position is that evaluators are best off when their positions are as secure and independent as possible from the influence of project management and staff. The other is that sustained contact with the policy and program staff enhances evaluators' work by providing a better understanding of the organization's objectives and activities while inspiring trust and thus increasing the evaluator's influence.

Second, there are ambiguities surrounding the role of the evaluator vis-à-vis program staff and groups of stakeholders regardless of whether the evaluator is an insider or outsider. The extent to which relations between staff members should resemble other structures in corporations or the collegial model that supposedly characterizes academia is an issue. But it is only one dimension to the challenge of structuring appropriate working relationships that confronts the evaluator.

Third, there is the concern on the part of evaluators with the "standing" of the organizations with which they are affiliated. Like universities, the settings in which evaluators work can be ranked and rated along a number of dimensions and a relatively few large evaluation organizations constitute a recognized elite subset of work places. Whether it is better to be a small fish in a big pond or vice versa is an issue in the evaluation field.

The discussion that follows, it bears emphasis, is based more on impressions of the authors of this text than on empirical research findings. Our impressions may be faulty, but it is a fact that debates surrounding these issues are commonplace whenever and wherever a critical mass of evaluators is found.

Inside Versus Outside Evaluations

In the past, some experienced evaluators went so far as to state categorically that evaluations should never be undertaken within the organization responsible for the administration of a project, but should always be conducted by an outside group. One reason "outsider" evaluations may have seemed the desired option is that there were differences in the levels of training and presumed competence of insider and outsider evaluation staffs. These differences have narrowed. The career of an evaluation researcher has typically taken one of three forms. Until the 1960s, a large proportion of evaluation research was done by either university-affiliated researchers or research firms. Since the late 1960s, public service agencies in various program areas have been hiring researchers for staff positions to conduct more in-house evaluations. Also, the proportion of evaluations done by private, for-profit research groups has increased markedly. As research

positions in both types of organizations have increased and the academic job market has declined, more persons who are well trained in the social and behavioral sciences have gravitated toward research jobs in public agencies and for-profit firms.

The current evidence is far from clear regarding whether inside or outside evaluations are more likely to be of higher technical quality. But technical quality is not the only criterion—utility may be just as important. A study in the Netherlands of external and internal evaluations suggests that internal evaluations may have a higher rate of impact on organizational decisions. According to van de Vall and Bolas (1981), of more importance than which category of researchers excels at social policy formation are those variables responsible for the higher rate of utilization of internal researchers' findings. The answer, they suggest, lies partly in a higher rate of communication between inside researchers and policymakers, accompanied by greater consensus, and partly in a balance between standards of epistemological and implemental validity: "In operational terms, this means that social policy researchers should seek equilibrium between time devoted to methodological perfection and translating results into policy measures" (p. 479). Their data suggest that currently in-house social researchers are in a more favorable position than external researchers for achieving these instrumental goals.

Given the increased competence of staff and the visibility and scrutiny of the evaluation enterprise, there is no reason now to favor one organizational arrangement over another. Nevertheless, there remain many critical points during an evaluation when there are opportunities for work to be misdirected and consequently misused irrespective of the locus of the evaluators. The important issue, there-

fore, is that any evaluation strikes an appropriate balance between technical quality and utility for its purposes, recognizing that those purposes may often be different for internal evaluations than for external ones.

Organizational Roles

Whether evaluators are insiders or outsiders, they need to cultivate clear understandings of their roles with sponsors and program staff. Evaluators' full comprehension of their roles and responsibilities is one major element in the successful conduct of an evaluation effort.

Again, the heterogeneity of the field makes it difficult to generalize on the best ways to develop and maintain the appropriate working relations.

One common mechanism is to have in place advisory groups or one or more consultants to oversee evaluations and provide some aura of authenticity to the findings. The ways such advisory groups or consultants work depend on whether an inside or an outside evaluation is involved and on the sophistication of both the evaluator and the program staff. For example, large-scale evaluations undertaken by federal agencies and major foundations often have advisory groups that meet regularly and assess the quality, quantity, and direction of the work. Some public and private health and welfare organizations with small evaluation units have consultants who provide technical advice to the evaluators or advise agency directors on the appropriateness of the evaluation units' activities, or both. Sometimes advisory groups and consultants are mere window dressing; we do not recommend their use if that is their only function. When members are actively engaged, however, advisory groups can be particularly useful in fostering interdisciplinary evaluation approaches, in adjudicating disputes between

program and evaluation staffs, and in defending evaluation findings in the face of concerted attacks by those whose interests are threatened.

EVALUATION STANDARDS, GUIDELINES, AND ETHICS

As the field of evaluation became increasingly professionalized, many evaluators began to pressure their professional associations to formulate and publish standards that could guide them in their evaluation work and in negotiations with evaluation funders and other major stakeholders. For example, it would be useful to evaluators to be able to bolster an argument for the right to freely publish evaluation findings if they could cite a published set of practice standards that included publication rights as standard evaluation practice. In addition, almost every practicing evaluator encounters situations requiring ethical judgments. For example, does an evaluator studying a child abuse prevention program have an obligation to report his observation of child abuse in a family revealed in the course of an interview on parenting practices? Published standards or practice guidelines also provide legitimacy to those who advertise their services as practices in conformity with them.

Two major efforts have been made to provide guidance to evaluators. Under the aegis of the American National Standards Institute (ANSI), the Joint Committee on Standards for Educational Evaluation (1994) has published *The Program Evaluation Standards*, now in its second edition. The Joint Committee is made up of representatives from several professional associations, including, among others, the American Evaluation Association, the Ameri-

can Psychological Association, and the American Educational Research Association. Originally set up to deal primarily with educational programs, the Joint Committee expanded its coverage to include all kinds of program evaluation. The *Standards* cover a wide variety of topics ranging from what provisions should appear in evaluation contracts through issues in dealing with human subjects to the standards for the analysis of quantitative and qualitative data. Each of the several score standards is illustrated with cases illustrating how the standards can be applied in specific instances.

The second major effort, *Guiding Principles for Evaluators* (Shadish, Newman, et al., 1995), has been adopted by the American Evaluation Association. Rather than proclaim standard practices, the *Guiding Principles* sets out five principles, quite general in character, for the guidance of evaluators. The principles follow, and the full statements are presented in Exhibit 12-E.

- A. *Systematic inquiry*: Evaluators conduct systematic, data-based inquiries about whatever is being evaluated.
- B. *Competence*: Evaluators provide competent performance to stakeholders.
- C. *Integrity/honesty*: Evaluators ensure the honesty and integrity of the entire evaluation process.
- D. *Respect for people*: Evaluators respect the security, dignity, and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.
- E. *Responsibilities for general and public welfare*: Evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.

These five principles are elaborated and discussed in the *Guiding Principles*, although not to the detailed extent to that found in the Joint Committee's work. Just how useful such general principles may be is problematic. An evaluator who has a specific ethical problem will likely find very little guidance in any one of the general principles. (See Shadish, Newman, et al., 1995, for critical appraisals of the *Guiding Principles*.)

We expect that developing a set of practice standards and ethical principles that can provide pointed advice to evaluators will take some time. The diversity of evaluation styles will make it difficult to adopt standards because any practice so designated may contradict what some group may consider good practice. The development of standards would be considerably advanced by the existence of "case law," the accumulation of adjudicated specific instances in which the principles have been applied. However, neither the Joint Committee's *Standards* nor the American Evaluation Association's *Guiding Principles* have any mode of enforcement, the usual institutional mechanism for the development of case law.

Until such evaluation standards and ethical rules are established, evaluators will have to rely on such general principles as the profession appears to be currently willing to endorse. A useful discussion of the many issues of applied ethics for program evaluation can be found in Newman and Brown (1996).


The Leadership Role of Evaluation "Elite" Organizations

A small group of evaluators, numbering perhaps no more than 1,000, constitutes an "elite" in the field by virtue of the scale of the evaluations they conduct and the size of the organizations for which they work. They are

EXHIBIT 12-E The American Evaluation Association's Guiding Principles for Evaluators


- A. *Systematic inquiry*: Evaluators conduct systematic, data-based inquiries about whatever is being evaluated.
 1. Evaluators should adhere to the highest appropriate technical standards in conducting their work, whether that work is quantitative or qualitative in nature, so as to increase the accuracy and credibility of the evaluative information they produce.
 2. Evaluators should explore with the client the shortcomings and strengths both of the various evaluation questions it might be productive to ask and the various approaches that might be used for answering those questions.
 3. When presenting their work, evaluators should communicate their methods and approaches accurately and in sufficient detail to allow others to understand, interpret, and critique their work. They should make clear the limitations of an evaluation and its results. Evaluators should discuss in a contextually appropriate way those values, assumptions, theories, methods, results, and analyses that significantly affect the interpretation of the evaluative findings. These statements apply to all aspects of the evaluation, from its initial conceptualization to the eventual use of findings.
- B. *Competence*: Evaluators provide competent performance to stakeholders.
 1. Evaluators should possess (or, here and elsewhere as appropriate, ensure that the evaluation team possesses) the education, abilities, skills, and experience appropriate to undertake the tasks proposed in the evaluation.
 2. Evaluators should practice within the limits of their professional training and competence and should decline to conduct evaluations that fall substantially outside those limits. When declining the commission or request is not feasible or appropriate, evaluators should make clear any significant limitations on the evaluation that might result. Evaluators should make every effort to gain the competence directly or through the assistance of others who possess the required expertise.
 3. Evaluators should continually seek to maintain and improve their competencies, in order to provide the highest level of performance in their evaluations. This continuing professional development might include formal coursework and workshops, self-study, evaluations of one's own practice, and working with other evaluators to learn from their skills and expertise.

(continued)

 **EXHIBIT 12-E** Continued

C. Integrity/honesty: Evaluators ensure the honesty and integrity of the entire evaluation process.

1. Evaluators should negotiate honestly with clients and relevant stakeholders concerning the costs, tasks to be undertaken, limitations of methodology, scope of results likely to be obtained, and uses of data resulting from a specific evaluation. It is primarily the evaluator's responsibility to initiate discussion and clarification of these matters, not the client's.
2. Evaluators should record all changes made in the originally negotiated project plans, and the reasons why the changes were made. If those changes would significantly affect the scope and likely results of the evaluation, the evaluator should inform the client and other important stakeholders in a timely fashion (barring good reason to the contrary, before proceeding with further work) of the changes and their likely impact.
3. Evaluators should seek to determine, and where appropriate be explicit about, their own, their clients', and other stakeholders' interests concerning the conduct and outcomes of an evaluation (including financial, political, and career interests).
4. Evaluators should disclose any roles or relationships they have concerning whatever is being evaluated that might pose a significant conflict of interest with their role as an evaluator. Any such conflict should be mentioned in reports of the evaluation results.
5. Evaluators should not misrepresent their procedures, data, or findings. Within reasonable limits, they should attempt to prevent or correct any substantial misuses of their work by others.
6. If evaluators determine that certain procedures or activities seem likely to produce misleading evaluative information or conclusions, they have the responsibility to communicate their concerns, and the reasons for them, to the client (the one who funds or requests the evaluation). If discussions with the client do not resolve these concerns, so that a misleading evaluation is then implemented, the evaluator may legitimately decline to conduct the evaluation if that is feasible and appropriate. If not, the evaluator should consult colleagues or relevant stakeholders about other proper ways to proceed (options might include, but are not limited to, discussions at a higher level, a dissenting cover letter or appendix, or refusal to sign the final document).
7. Barring compelling reason to the contrary, evaluators should disclose all sources of financial support for an evaluation, and the source of the request for the evaluation.

 **EXHIBIT 12-E** Continued

D. Respect for people: Evaluators respect the security, dignity, and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.

1. Where applicable, evaluators must abide by current professional ethics and standards regarding risks, harms, and burdens that might be engendered to those participating in the evaluation; regarding informed consent for participation in evaluation; and regarding informing participants about the scope and limits of confidentiality. Examples of such standards include federal regulations about protection of human subjects, or the ethical principles of such associations as the American Anthropological Association, the American Educational Research Association, or the American Psychological Association. Although this principle is not intended to extend the applicability of such ethics and standards beyond their current scope, evaluators should abide by them where it is feasible and desirable to do so.
2. Because justified negative or critical conclusions from an evaluation must be explicitly stated, evaluations sometimes produce results that harm client or stakeholder interests. Under this circumstance, evaluators should seek to maximize the benefits and reduce any unnecessary harms that might occur, provided this will not compromise the integrity of the evaluation findings. Evaluators should carefully judge when the benefits from doing the evaluation or in performing certain evaluation procedures should be forgone because of the risks or harms. Where possible, these issues should be anticipated during the negotiation of the evaluation.
3. Knowing that evaluations often will negatively affect the interests of some stakeholders, evaluators should conduct the evaluation and communicate its results in a way that clearly respects the stakeholders' dignity and self-worth.
4. Where feasible, evaluators should attempt to foster the social equity of the evaluation, so that those who give to the evaluation can receive some benefits in return. For example, evaluators should seek to ensure that those who bear the burdens of contributing data and incurring any risks are doing so willingly and that they have full knowledge of, and maximum feasible opportunity to obtain, any benefits that may be produced from the evaluation. When it would not endanger the integrity of the evaluation, respondents or program participants should be informed if and how they can receive services to which they are otherwise entitled without participating in the evaluation.
5. Evaluators have the responsibility to identify and respect differences among participants, such as differences in their culture, religion, gender, disability, age, sexual orientation, and ethnicity, and to be mindful of potential implications of these differences when planning, conducting, analyzing, and reporting their evaluations.

(continued)

 **EXHIBIT 12-E** Continued

E. *Responsibilities for general and public welfare: Evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.*

1. When planning and reporting evaluations, evaluators should consider including important perspectives and interests of the full range of stakeholders in the object being evaluated. Evaluators should carefully consider the justification when omitting important value perspectives or the views of important groups.
2. Evaluators should consider not only the immediate operations and outcomes of whatever is being evaluated but also the broad assumptions, implications, and potential side effects of it.
3. Freedom of information is essential in a democracy. Hence, barring compelling reason to the contrary, evaluators should allow all relevant stakeholders to have access to evaluative information and should actively disseminate that information to stakeholders if resources allow. If different evaluation results are communicated in forms that are tailored to the interests of different stakeholders, those communications should ensure that each stakeholder group is aware of the existence of the other communications. Communications that are tailored to a given stakeholder should always include all important results that may bear on interests of that stakeholder. In all cases, evaluators should strive to present results as clearly and simply as accuracy allows so that clients and other stakeholders can easily understand the evaluation process and results.
4. Evaluators should maintain a balance between client needs and other needs. Evaluators necessarily have a special relationship with the client who funds or requests the evaluation. By virtue of that relationship, evaluators must strive to meet legitimate client needs whenever it is feasible and appropriate to do so. However, that relationship can also place evaluators in difficult dilemmas when client interests conflict with other interests, or when client interests conflict with the obligation of evaluators for systematic inquiry, competence, integrity, and respect for people. In these cases, evaluators should explicitly identify and discuss the conflicts with the client and relevant stakeholders, resolve them when possible, determine whether continued work on the evaluation is advisable if the conflicts cannot be resolved, and make clear any significant limitations on the evaluation that might result if the conflict is not resolved.
5. Evaluators have obligations that encompass the public interest and good. These obligations are especially important when evaluators are supported by publicly generated funds, but clear threats to the public good should never be ignored in any evaluation. Because the public interest and good are rarely the same as the interests of any particular group (including those of the client or funding agency), evaluators will usually have to go beyond an analysis of particular stakeholder interests when considering the welfare of society as a whole.

SOURCE: American Evaluation Association, Task Force on Guiding Principles for Evaluators, *Guiding Principles for Evaluators*, New Directions for Evaluation, no. 66 (San Francisco: Jossey-Bass, 1995), pp. 19-26.

somewhat akin to those elite physicians who practice in the hospitals of important medical schools. They and their settings are few in number but powerful in establishing the norms for the field; the ways in which they work and the standards of performance in their organizations represent an important version of professionalism that persons in other settings may use as a role model.

The number of organizations that carry out national or otherwise large-scale evaluations with a high degree of technical competence is quite small. But in terms of both visibility and evaluation dollars expended, these organizations occupy a strategic position in the field. Most of the large federal evaluation contracts over the years have been awarded to a small group of profit-making research firms (such as Abt Associates, Mathematica Policy Research, and Westat, to name a few) and not-for-profit research organizations and universities (examples are Battelle Memorial Institute, the RAND Corporation, the Research Triangle Institute, the Urban Institute, and the Manpower Development Research Corporation). A handful of research-oriented universities with affiliated research institutes—the National Opinion Research Center (NORC) at the University of Chicago, the Institute for Research on Poverty at the University of Wisconsin (until recently), and the Institute for Social Research at the University of Michigan, for example—also receive grants and contracts for undertaking large-scale evaluations. In addition, the evaluation units of federal agencies that contract for and fund evaluation research, and a few of the large national foundations, include significant numbers of highly trained evaluators on their staffs. Within the federal government, perhaps the highest concentration of skilled evaluators was to be found until recently in the Program Evaluation and Methodology Division of the

GAO, where a large group of evaluation specialists has extended the activities of this key “watchdog” organization from auditing to assessing appropriate program implementation and estimating the impact of federal initiatives.

One of the features of these elite for-profit and nonprofit organizations that are the contractors for most large-scale evaluations is a continual concern with the quality of their work. In part, this has come about because of earlier critiques of their efforts, which formerly were not as well conducted technically as those done by persons in academic institutions (Bernstein and Freeman, 1975). But as they came to dominate the field, at least in terms of large-scale evaluations, and as they found sponsors of evaluations increasingly using criteria of technical competence in selecting contractors, their efforts improved markedly from a methodological standpoint. So, too, have the competencies of their staffs, and they now compete for the best-trained persons in applied work. Also, they have found it to be in their self-interest to encourage staff to publish in professional journals, participate actively in professional organizations, and engage in frontier efforts to improve the state of the art. To the extent that there is a general movement toward professionalism, these organizations are its leaders.

UTILIZATION OF EVALUATION RESULTS

In the end, the worth of evaluations must be judged by their utility. For this reason, considerable thought and research have been devoted to the use of evaluation results. As a starting point, the conventional three-way classification of the ways evaluations are used is

helpful (Leviton and Hughes, 1981; Rich, 1977; Weiss, 1988).

First, evaluators prize the *direct* or *instrumental* use of their evaluations. By direct use is meant the documented and specific use of evaluation findings by decisionmakers and other stakeholders. For example, evaluators' data showing that patients of health maintenance organizations are hospitalized fewer days than patients who are treated in the ambulatory clinics of hospitals have been used by Congress and health policymakers in developing medical care programs for the poor (Freeman, Kiecolt, and Allen, 1982). More recently, the excellent field experiments conducted by the Manpower Development Research Corporation on workfare conducted under AFDC waivers (Gueron and Pauly, 1991) are influencing how the states are currently reforming welfare.

Second, utilization can be conceptual. As Rich (1977) defined it, *conceptual utilization* refers to the use of evaluations to influence thinking about issues in a general way. An example is the current effort to control the costs of delivering health and welfare services, stimulated at least in part by evaluations of the efficacy of these services and their costs-to-benefits ratios. These evaluations did not lead to the adoption of specific programs or policies but provided evidence that present ways of delivering health care were costly and inefficient.

Third, *persuasive utilization* refers to enlisting evaluation results in efforts either to support or to refute political positions—in other words, to defend or attack the status quo. For example, one of the frequent rationales used by the Reagan administration in defending the cutting of social programs was the lack of clear findings of positive impact in the evaluations of major programs. Persuasive use is

similar to speechwriters' inserting quotes into political speeches, whether they are applicable or not. For the most part, the persuasive use of evaluations is out of the hands of program evaluators and sponsors alike and will not concern us further.

Do Evaluations Have Direct Utility?

Disappointment about the extent of the utilization of evaluations apparently is due to their limited direct or instrumental use. It is clear that many evaluations initiated for their direct utility fell short of that mark. However, it is only in the past decade that the extent of direct use has been systematically studied. These recent efforts challenge the previously held belief that evaluations do not have direct utility.

One careful study (Leviton and Boruch, 1983), for example, examined the direct use of evaluations sponsored by the U.S. Department of Education. They found numerous instances in which the results of evaluations led to important program changes and even more incidents in which they were influential inputs, though not the sole inputs, in the decision-making process.

Chelimsky (1991) also cites several instances in which social science research provided critical knowledge for the development of public policy. Unfortunately, large-scale evaluations typically dominate the printed literature. The many small-scale evaluations, especially those that are diagnostic and formative, that have experienced direct use in improving programs do not ordinarily find their way into the printed literature.

Nevertheless, contrary to the views expressed in earlier editions of this book, there does seem to be a fair degree of instrumental

utilization, although a pessimistic view on this point is still widely held among both evaluators and potential consumers of evaluations.

Subsequently, we will suggest means to increase the utilization of evaluations. Most of these suggestions are particularly relevant to increasing the direct use of studies. However, it is also important to appropriately value the conceptual use of evaluations.

Conceptual Use of Evaluations

No doubt every evaluator has had moments of glorious dreams in which a grateful world receives with adulation the findings of his or her evaluation and puts the results immediately and directly to use. Most of our dreams must remain dreams. We would argue, however, that the conceptual use of evaluations often provides important inputs into policy and program development and should not be compared with finishing the race in second place. Conceptual utilization may not be as visible to peers or sponsors, yet this use of evaluations deeply affects the community as a whole or critical segments of it.

By *conceptual use* we refer to the variety of ways in which evaluations indirectly have an impact on policies, programs, and procedures. This impact ranges from sensitizing persons and groups to current and emerging social problems to influencing future program and policy development by contributing to the cumulative results of a series of evaluations.

Evaluations perform a sensitizing role by documenting the incidence, prevalence, and distinguishing features of social problems. Diagnostic evaluation activities, described in Chapter 4, have provided clearer and more precise understanding of changes occurring in the family system, critical information on the

location and distribution of unemployed persons, and other meaningful descriptions of the social world.

Impact assessments, too, have conceptual utility. A specific example is the current concern with "notch" groups in the development of medical care policy. Evaluations of programs to provide medical care to the poor have found that the very poor, those who are eligible for public programs such as Medicaid, often are adequately provided with health services. Those just above them—the "notch" group—who are not eligible for public programs tend to fall in the cracks between public assistance and being able to provide for their own care. They have decidedly more difficulty receiving services, and, when seriously ill, represent a major burden on community hospitals, which cannot turn them away yet can receive reimbursement neither from the patients nor from the government. Concern with the near-poor, or notch group, is increasing because of their exclusion from a wide range of health, mental health, and social service programs.

An interesting example of a study that had considerable long-term impact is the now classic Coleman report on educational opportunity (Coleman et al., 1966). The initial impetus for this study was a 1964 congressional mandate to the (then) Office of Education to provide information on the quality of educational opportunities provided to minority students in the United States. Its actual effect was much more far-reaching: The report changed the conventional wisdom about the characteristics of good and bad educational settings, turning policy and program interest away from problems of fiscal support to ways of improving teaching methods (Moynihan, 1991).

The conceptual use of evaluation results creeps into the policy and program worlds by a variety of routes, usually circuitous, that are

difficult to trace. For example, Coleman's report to the Office of Education did not become a Government Printing Office best-seller. It is unlikely that more than a few hundred people actually read it cover to cover. But journalists wrote about it, essayists summarized its arguments, and major editorial writers mentioned it. Through these communication brokers, the findings became known to policymakers in the education field and to politicians at all levels of government.

In 1967, a year after his report had been published by the Government Printing Office, Coleman was convinced that it had been buried in the National Archives and would never emerge again. Eventually, however, his findings in one form or another reached a wide and influential audience. Indeed, by the time Caplan and his associates (Caplan and Nelson, 1973) questioned influential political figures in Washington about which social scientists had influenced them, Coleman's name was among the most prominently and consistently mentioned.

Some of the conceptual utilizations of evaluations may be described simply as consciousness-raising. For example, the development of early-childhood education programs was stimulated by the evaluation findings resulting from an impact assessment of *Sesame Street*. The evaluation found that although the program did have an effect on young children's educational skills, the magnitude of the effect was not as large as the program staff and sponsors imagined it would be. Prior to the evaluation, some educators were convinced that the program represented the "ultimate" solution and that they could turn their attention to other educational problems. The evaluation findings led to the conviction that early-child-

hood education was in need of further research and development.

As in the case of direct utilization, evaluators have an obligation to do their work in ways that maximize conceptual utilization. In a sense, efforts at maximizing conceptual utilization are more difficult to devise than ones to optimize direct use. To the extent that evaluators are hired guns and turn to new ventures after completing an evaluation, they may not be around or have the resources to follow through on promoting conceptual utilization. Sponsors of evaluations and other stakeholders who more consistently maintain a commitment to particular social policy and social problem areas must assume at least some of the responsibility, if not the major portion, for maximizing the conceptual use of evaluations. Often these parties are in a position to perform the broker function alluded to earlier.

Variables Affecting Utilization

In studies of the use of social research in general, and evaluations in particular, five conditions appear to affect utilization consistently (Leviton and Hughes, 1981):

- Relevance
- Communication between researchers and users
- Information processing by users
- Plausibility of research results
- User involvement or advocacy

The importance of these conditions and their relative contributions to utilization have been carefully studied by Weiss and Bucuvalas (1980). They examined 155 decisionmakers in the mental health field and their reactions to

EXHIBIT 12-F Truth Tests and Utility Tests

In coping with incoming floods of information, decisionmakers invoke three basic frames of reference. One is the relevance of the content of the study to their sphere of responsibility, another is the trustworthiness of the study, and the third is the direction that it provides. The latter two frames, which we have called truth and utility tests, are each composed of two interdependent components:

Truth tests—Is the research trustworthy? Can I rely on it? Will it hold up under attack? The two specific components are

1. Research quality: Was the research conducted by proper scientific methods?
2. Conformity to user expectations: Are the results compatible with my experience, knowledge, and values?

Utility tests—Does the research provide direction? Does it yield guidance either for immediate action or for considering alternative approaches to problems? The two specific components are

1. Action orientation: Does the research show how to make feasible changes in things that can feasibly be changed?

SOURCE: Adapted, with permission, from C. H. Weiss and M. J. Bucuvalas, "Truth Tests and Utility Tests: Decision-Makers' Frames of Reference for Social Science Research," *American Sociological Review*, April 1980, 45:302-313.

2. Challenge to the status quo: Does the research challenge current philosophy, program, or practice? Does it offer new perspectives?

Together with relevance (i.e., the match between the topic of the research and the person's job responsibilities), the four components listed above constitute the frames of reference by which decisionmakers assess social science research. Research quality and conformity to user expectations form a single truth test in that their effects are contingent on each other: Research quality is less important for the usefulness of a study when results are congruent with officials' prior knowledge than when results are unexpected or counterintuitive. Action orientation and challenge to the status quo represent alternative functions that a study can serve. They constitute a utility test, since the kind of explicit and practical direction captured by the action orientation frame is more important for a study's usefulness when the study provides little criticism or reorientation (challenge to the status quo) than it is when challenge is high. Conversely, the criticisms of programs and the new perspectives embedded in challenge to the status quo add more to usefulness when a study lacks prescriptions for implementation.

50 actual research reports. Decisionmakers, they found, apply both a *truth test* and a *utility test* in screening social research reports. Truth is judged on two bases: research quality and conformity to prior knowledge and expecta-

tions. Utility refers to feasibility potential and the degree of challenge to current policy. The Weiss and Bucuvalas study provides convincing evidence of the complexity of the utilization process (see Exhibit 12-F).

Guidelines for Maximizing Utilization

Out of the research on utilization and the real-world experiences of evaluators, a number of guidelines for increasing utilization have emerged. These have been summarized by Solomon and Shortell (1981) and are briefly noted here for reference:

1. *Evaluators must understand the cognitive styles of decisionmakers.* For instance, there is no point in presenting a complex piece of analysis to a politician who cannot or will not consume such material. Thus, reports and oral presentations tailored to a predetermined audience may be more appropriate than, say, academic journal articles.

2. *Evaluation results must be timely and available when needed.* Evaluators must therefore balance thoroughness and completeness of analysis with timing and accessibility of findings. In doing so, they may have to risk criticism from some of their academic colleagues, whose concepts of scholarship cannot always be met because of the need for rapid results and crisp reporting.

3. *Evaluations must respect stakeholders' program commitments.* The usefulness of evaluations depends on wide participation in the evaluation design process to ensure sensitivity to various stakeholders' interests. Differences in values and outlooks between clients and evaluators should be explicated at the outset of a study and be a determinant of whether a specific evaluation is undertaken by a particular evaluation team.

4. *Utilization and dissemination plans should be part of the evaluation design.* Evaluation findings are most likely to be used if the

evaluation effort includes "teaching" potential users the strengths and limitations of the effort, the degree to which one may expect definitive results, how the information from the evaluation can be effectively communicated by decisionmakers to their constituencies, and what criticisms and other reactions may be anticipated.

5. *Evaluations should include an assessment of utilization.* Evaluators and decisionmakers must not only share an understanding of the purposes for which a study is undertaken but also agree on the criteria by which its successful utilization may be judged. Under such conditions, however much informality is necessary, an effort should be made to judge the extent to which the uses of findings meet these expectations.

Although these guidelines are relevant to the utilization of all program evaluations, the roles of evaluation consumers do differ. Clearly, these differing roles affect the uses to which information is put and, consequently, the choice of mechanisms for maximizing utility. For example, if evaluations are to influence federal legislation and policies, they must be conducted and "packaged" in ways that meet the needs of congressional staff. For the case of educational evaluation and legislation, Florio, Behrmann, and Goltz (1979) furnished a useful summary of requirements that rings as true today as when it was compiled (see Exhibit 12-G).

EPILOGUE

There are many reasons to expect continued support of evaluation activities. First, decision-

EXHIBIT 12-G Educational Evaluation: The Unmet Potential

The interviewees (congressional staff involved in developing educational legislation) mentioned more than 90 steps that could be taken to improve the use of educational studies in the formation of legislative policy. The most common themes, which reflect the current barriers to such use, are the ways in which research and assessment reports are presented and the failure to meet the needs demanded by the policy cycles in Congress. Staffers struck a common theme of work and information overload problems associated with the job. They rarely have time to evaluate the evaluations, let alone read through the voluminous reports that come across their desks. This was at the root of the repeated call for executive summaries in the front matter of reports, which would allow them to judge the relevance of the contents and determine whether further reading for substance was necessary. Although 16 (61%) of the staffers complained of an information overload problem, 19 also indicated that they were often forced to generate their own data relevant to political and policy questions. As one staffer put it, "We have no overload of *useful and understandable* information."

The timing of study reports and their relevance to questions before the Congress were major barriers repeatedly mentioned by congressional staff. A senior policy analyst for the Assistant Secretary of Education compared the policy process to a moving train. She suggested that information providers have the obligation to

know the policy cycle and meet it on its own terms. The credibility problem is also one that plagues social inquiry. The Deputy Director of the White House Domestic Policy staff said that all social science suffers from the perception that it is unreliable and not policy-relevant. His comments were reflected by several of the staffers interviewed; for example, "Research rarely provides definitive conclusions," or "For every finding, others negate it," or "Educational research can rarely be replicated and there are few standards that can be applied to assess the research products." One went so far as to call project evaluations lies, then reconsidered and called them embellishments.

It must be pointed out that the distinctions among different types of inquiry research, evaluation, data collection, and so on are rarely made by the recipients of knowledge and information. If project evaluations are viewed as fabrications, it reflects negatively on the entire educational inquiry community. Even when policy-relevant research is presented in time to meet the moving train, staffers complain of having too much information that cannot be easily assimilated, or that studies are poorly packaged, contain too much technical jargon, and are too self-serving. Several said that researchers write for other researchers and rarely, except in congressionally mandated studies, tailor their language to the decision-making audiences in the legislative process.

SOURCE: Adapted from D. H. Florio, M. M. Behrmann, and D. L. Goltz, "What Do Policy Makers Think of Evaluational Research and Evaluation? Or Do They?" *Educational Evaluation and Policy Analysis*, 1979, 1(6):61-87. Copyright 1979 by the American Educational Research Association, Washington, DC. Adapted by permission of the publisher.

makers, planners, project staffs, and target participants are increasingly skeptical of common sense and conventional wisdom as sufficient bases on which to design social programs that will achieve their intended goals. Decades of attempts to solve the problems represented by explosive population growth, the maldistribution of resources within and between societies, popular discontent, crime, educational deficiencies among adults and children, drug and alcohol abuse, and weaknesses in traditional institutions such as the family have led to a realization that these are obstinate and difficult issues. This skepticism has, in turn, led policymakers and decisionmakers to seek ways to learn more quickly and efficiently from their mistakes and to capitalize more rapidly on measures that work.

A second major reason for the growth of evaluation research has been the development of knowledge and technical procedures in the social sciences. The refinement of sample survey procedures has provided an important information-gathering method. When coupled with more traditional experimental methods in the form of field experiments, these procedures become a powerful means of testing social programs. Advances in measurement, statistical theory, and substantive knowledge in the social sciences have added to the ability of social scientists to take on the special tasks of evaluation research.

Finally, there are the changes in the social and political climate of our times. As a society—indeed, as a world—we have come to insist that communal and personal problems are not fixed features of the human condition but can be ameliorated through the reconstruction of social institutions. We believe more than our ancestors did that societies can be improved and that the lot of all persons can be enhanced by the betterment of the disadvantaged and deprived. At the same time, we are confronted with severely limited resources for welfare, health, and other social programs. It is tempting simply to wish away unemployment, crime, homelessness—all the social ills we are too familiar with—and to believe that “moral reconstruction” will diminish the need for effective and efficient social programs. But it is catastrophically naive to think that doing so will solve our problems.

The prognosis is troublesome, in the short term at least, when we contemplate both the variety and number of concerns that require urgent action and the level of resources being committed to controlling and ameliorating them. It is clear that sensible, orderly procedures are required to choose which problems to confront first, and which programs to implement to deal with them. Our position is clear: Systematic evaluations are invaluable to current and future efforts to improve the lot of humankind.

SUMMARY

- ✎ Evaluation is purposeful, applied social research. In contrast to basic research, evaluation is undertaken to solve practical problems. Its practitioners must be conversant with methods from several disciplines and able to apply them to many types of problems. Furthermore, the criteria for judging the work include its utilization and hence its impact on programs and the human condition.
- ✎ Evaluators must put a high priority on deliberately planning for the dissemination of the results of their work. In particular, they need to become “secondary disseminators” who package their findings in ways that are geared to the needs and competencies of a broad range of relevant stakeholders.
- ✎ Because the value of their work depends on its utilization by others, evaluators must understand the social ecology of the arena in which they work.
- ✎ Evaluation is directed to a range of stakeholders with varying and sometimes conflicting needs, interests, and perspectives. Evaluators must determine the perspective from which a given evaluation should be conducted, explicitly acknowledge the existence of other perspectives, be prepared for criticism even from the sponsors of the evaluation, and adjust their communication to the requirements of various stakeholders.
- ✎ An evaluation is only one ingredient in a political process of balancing interests and coming to decisions. The evaluator’s role is close to that of an expert witness, furnishing the best information possible under the circumstances; it is not the role of judge and jury.
- ✎ Two significant strains that result from the political nature of evaluation are (a) the different requirements of political time and evaluation time, and (b) the need for evaluations to have policy-making relevance and significance. With respect to both of these sets of issues, evaluators must look beyond considerations of technical excellence and pure science, mindful of the larger context in which they are working and the purposes being served by the evaluation.
- ✎ Evaluators are perhaps better described as a “near-group” than as a profession. The field is marked by diversity in disciplinary training, type of schooling, perspectives on appropriate methods, and an absence of strong communication among its practitioners. Although the field’s rich diversity is one of its attractions, it also leads to unevenness in competency, lack of consensus on appropriate approaches, and justifiable criticism of the methods used by some evaluators.
- ✎ Among the enduring controversies in the field has been the issue of qualitative and quantitative research. Stated in the abstract, the issue is a false one; the two approaches are suitable for different and complementary purposes.

- ✎ Evaluators are also diverse in their activities and working arrangements. Although there has been considerable debate over whether evaluators should be independent of program staff, there is now little reason to prefer either inside or outside evaluation categorically. What is crucial is that evaluators have a clear understanding of their role in a given situation.
- ✎ There is reason to be concerned about the field's being dominated by a small group of elite evaluation organizations and their staffs. Although these organizations contribute to the movement toward professionalization of the field, efforts to enhance opportunities for career mobility and interaction in the profession are desirable.
- ✎ Evaluative studies are worthwhile only if they are used. Three types of utilization are direct, or instrumental; conceptual; and persuasive. Although in the past, considerable doubt has been shed on the direct utility of evaluations, there is reason to believe they do have an impact on program development and modification. At least as important, the conceptual utilization of evaluations appears to have a definite effect on policy and program development, as well as social priorities, albeit one that is not always easy to trace.

GLOSSARY

Accessibility	The extent to which the structural and organizational arrangements facilitate participation in the program.
Accountability	The responsibility of program staff to provide evidence to stakeholders and sponsors that a program is effective and in conformity with its coverage, service, legal, and fiscal requirements.
Accounting perspectives	Perspectives underlying decisions on which categories of goods and services to include as costs or benefits in an analysis.
Administrative standards	Stipulated achievement levels set by program administrators or other responsible parties, for example, intake for 90% of the referrals within one month. These levels may be set on the basis of past experience, the performance of comparable programs, or professional judgment.
Articulated program theory	An explicitly stated version of program theory that is spelled out in some detail as part of a program's documentation and identity or as a result of efforts by the evaluator and stakeholders to formulate the theory.
Assessment of program process	An evaluative study that answers questions about program operations, implementation, and service delivery. Also known as a process evaluation or an implementation assessment.
Assessment of program theory	An evaluative study that answers questions about the conceptualization and design of a program.
Benefits	Net program outcomes, usually translated into monetary terms. Benefits may include both direct and indirect effects.
Benefits-to-costs ratio	The total discounted benefits divided by the total discounted costs.
Bias in coverage	The extent to which subgroups of a target population participate differentially in a program.
Black box evaluation	Evaluation of program outcomes without the benefit of an articulated program theory to provide insight into what is presumed to be causing those outcomes and why.