

Evaluation Strategies for Human Services Programs

A Guide for Policymakers and Providers

Adele Harrell
with
Martha Burt
Harry Hatry
Shelli Rossman
Jeffrey Roth
William Sabol

The Urban Institute
Washington, D.C.

Contents

Clarifying the Evaluation Questions, 2
Developing a Logic Model, 3
Assessing Readiness for Evaluation, 7
Selecting an Evaluation Design, 8
Identifying Potential Evaluation Problems, 25
Conclusions, 28

EXHIBITS

Exhibit A: Logic Model Used in Evaluation of the Children At Risk Program, 6
Exhibit B: Process for Selecting Impact Evaluation Designs, 17

Evaluation Strategies for Human Services Programs

A Guide for Policymakers and Providers

In the continuing effort to improve human service programs, funders, policymakers, and service providers are increasingly recognizing the importance of rigorous program evaluations. They want to know what the programs accomplish, what they cost, and how they should be operated to achieve maximum cost-effectiveness. They want to know which programs work for which groups, and they want conclusions based on evidence, rather than testimonials and impassioned pleas.

This paper lays out, for the non-technician, the basic principles of program evaluation design. It signals common pitfalls, identifies constraints that need to be considered, and presents ideas for solving potential problems. These principles are general and can be applied to a wide range of human service programs. We illustrate these principles here with examples from programs for vulnerable children and youth. Evaluation of these programs is particularly challenging because they address a wide diversity of problems and possible solutions, often include multiple agencies and clients, and change over time to meet shifting service needs.

Steps in Selecting the Appropriate Evaluation Design. The first step in the process of selecting an evaluation design is to clarify the questions that need to be answered. The next step is to develop a logic model that lays out the expected causal linkages between the program (or program components) and the program goals. Without tracing these anticipated links it is impossible to interpret the evaluation evidence that is collected. The third step is to review the program to assess its readiness for evaluation. These three steps can be done at the same time or in overlapping stages. For expositional clarity we will discuss each of them in turn. We will then describe how to select the best design for a given purpose from among the major types of evaluation that exist.

Clarifying the Evaluation Questions

The design of any evaluation begins by defining the audience for the evaluation findings, what they need to know, and when. These questions determine which of the following four major types of evaluation should be chosen:

Impact evaluations focus on questions of causality. Did the program have its intended effects? If so, who was helped and what activities or characteristics of the program created the impact? Did the program have any unintended consequences, positive or negative?

Performance monitoring provides information on key aspects of how a system or program is operating and the extent to which specified program objectives are being attained (e.g., numbers of youth served compared to target goals, reductions in school dropouts compared to target

goals). Results are used by service providers, funders, and policymakers to assess the program's performance and accomplishments.

Process evaluations answer questions about how the program operates and document the procedures and activities undertaken in service delivery. Such evaluations help identify problems faced in delivering services and strategies for overcoming these problems. They are useful to practitioners and service providers in replicating or adapting program strategies.

Cost evaluations address how much the program or program components cost, preferably in relation to alternative uses of the same resources and to the benefits being produced by the program. In the current fiscal environment, programs must expect to defend their costs against alternative uses.

A comprehensive evaluation will include all these activities. Sometimes, however, the questions raised, the target audience for findings, or the available resources limit the evaluation focus to one or two of these activities.

Whether to provide preliminary evaluations to staff for use in improving program operations and developing additional services is an issue that needs to be faced. Preliminary results can be effectively used to identify operational problems and develop the capacity of program staff to conduct their own ongoing evaluation and monitoring activities.(1) But this use of evaluation findings, called formative evaluations, presents a challenge to evaluators who are faced with the much more difficult task of estimating the impact of an evolving intervention. When the program itself is continuing to change, measuring impact requires ongoing measurement of the types and level of service provided. The danger in formative evaluations is that the line between program operations and assessment will be blurred. The extra effort and resources required for impact analysis in formative evaluations has to be measured against the potential gains to the program from ongoing improvements and the greater usefulness of the final evaluation findings.

Developing a Logic Model

It is impossible to interpret evaluation findings without a clear understanding of program goals, implementation sequences, and the expected links between them and expected program benefits. Expectations about these linkages are made explicit by developing a logic model. Such a model is developed by discussing with service providers and funders the goals of and rationales behind program organization and content, examining planning documents and program reports, and reviewing research findings on similar programs or problems. The literature review may be particularly helpful in identifying plausible causal links and any factors other than the program which should be considered in the evaluation.

The logic model provides a simplified description of the **program**, the intended **outputs**, and the intended **outcomes**. Program characteristics include the population to be reached, the resources to be used, and identification of the types and levels of service elements. Outputs are immediate program products resulting from the internal operations of the program, such as the delivery of planned services. Examples of output indicators in the area of programs for vulnerable children

and youth might include the numbers of children immunized, home visits by case managers, or youth completing a job training program. These program outputs are, in turn, the vehicle for producing the desired program outcomes, for example, decreases in childhood illnesses, decreases in abuse and neglect cases, or increases in youth employment. Careful attention must be paid to when the anticipated outcome should be expected to occur. For this reason it is often useful to divide outcomes into intermediate versus longer term. For example, improved school attendance in early grades might be an intermediate outcome associated with the longer-term outcome of dropout prevention. Care must be given to focusing on outcomes which will occur within the study period.

A classic failure in selecting an outcome that is expected to occur within the time frame of the study occurred in evaluations of the DARE drug prevention program, an educational program for fifth and sixth graders designed to prevent drug use. Evaluation results showed no significant prevention of drug use at the end of the program. This result should have been anticipated, since drug use does not typically begin among youth in this country until the mid-teen years (14 to 17). An age-appropriate intermediate outcome should have been selected as the primary outcome measure, such as improved peer resistance skills and changes in beliefs about the risks of drug use.

The logic model should also include explicit mapping of the conditions present in the program environment or characteristics of the target group or community that may affect the program's ability to achieve its goals. Non-program characteristics of the program organization, community or target population that are likely to influence the outputs and outcomes and/or use of program services are called **antecedent variables**. Conditions or events in the program, target population, or community that may limit or expand the extent to which program outputs actually produce the desired outcomes are called **mediating variables**. For example, a drug abuse prevention program may be less effective if the program staff are inexperienced, or if the local community offers fewer recreational alternatives to substance abuse and/or more active open drug markets (antecedent variables). Offering other support services in combination with the program may enhance its impact (a mediating variable).

In impact evaluations the logic model is used to spell out how, and for whom, certain services are expected to create specific changes/benefits. For example, if the program includes parenting classes, the logic model will identify this activity as a key program component and show the types of changes in parenting that will be used to measure program outcomes (e.g., by improving parental assistance with homework or helping parents communicate more effectively with adolescents).

In performance monitoring, the logic model is used to focus on which kinds of output and outcome indicators are appropriate for specific target populations, communities, or time periods. For example, among indicators of child improvement in school, one might expect attendance to improve in the first semester of a program, but academic test score improvement only after a significant period of program participation-with the timing possibly varying by the age and developmental stage of the children.

In process evaluation, the logic model is used to identify expectations about how the program should work-an "ideal type"-which can then be used to assess the deviations in practice, why these deviations have occurred, and how the deviations may affect program outputs. This assists program managers (and evaluators) to identify differences (including positive and negative unintended consequences), consider possible mechanisms for fine-tuning program operations to align the actual program with the planned approach, or re-visit program strategies to consider alternatives.(2)

Logic models are constructed to show temporal sequences, building left to right, and they typically diagram relationships with arrows. An example of a logic model is shown in Exhibit A. It was developed by the Urban Institute during the planning of the evaluation of the Children At Risk program (CAR). CAR is an intensive intervention program designed to prevent involvement in drugs and crime, and to foster healthy development among adolescents ages 13 to 15 who exhibit serious risk indicators and live in severely distressed inner-city neighborhoods.

The intervention consists of eight required program components:

Case Managers employed by the program make a service plan for all members of the household of participating youth and provide intensive follow-up on referrals to needed services, handling a caseload of 15;

Family Services include parenting skills training for all parents, and referral to other services as needed (intensive family counseling, stress management/coping skills training, identification and treatment of substance abuse, health care, job training and employment programs, housing, and income support services);

Education Services include tutoring or homework assistance for all youth, and referral to other services as needed (educational testing, special education classes);

After-School and Summer Activities for all CAR youth include recreational programs and life-skill/leadership development activities, combined with training or education;

Mentoring is provided by local organizations for youth in need of a caring relationship with an adult. The role of the mentor is to: (a) inform youth about alternative available choices (e.g., activities and goals); (b) familiarize them with strategies available for pursuing those choices; (c) provide training, opportunities for practice, and feedback in the development of skills for implementing particular strategies; and (d) provide relationships through which youth are affirmed, inspired, and encouraged to make healthy choices;

Incentives such as gifts and special events are used to build morale and attachment to the pro-social goals of the program (e.g., gift certificates, trips, and vouchers for pizza, sports shops, movies, and stipends for community service during summer programs);

Community Policing/Enhanced Enforcement is used in all target neighborhoods to create safer environments with less drug activity. Law enforcement activities include out-stationing

police in schools and neighborhood locations to maintain order and enhance relationships with community groups;

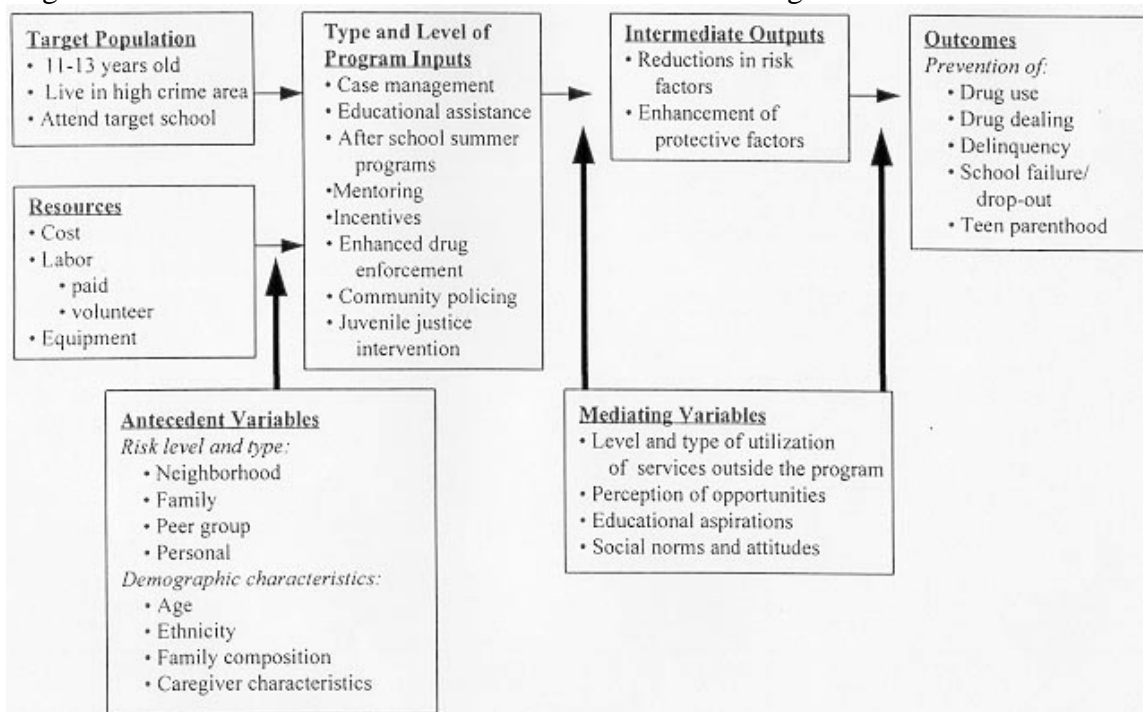
Criminal/juvenile Justice Intervention involves collaboration between case managers and juvenile court personnel to provide community service opportunities and enhanced supervision of youth in the justice system.

Antecedent variables include the levels and types of neighborhood, family, peer group, and personal risk factors for participants as well as their demographic characteristics. These are influences that are present before the program intervention.

Mediating variables include exposure to other social or educational services, perceptions of opportunities, and social norms. These are influences that operate at the same time as the program is operating. The program components are designed to achieve the intermediate outcomes—reductions in risk factors and enhancement of protective factors at the end of program participation.

Exhibit A

Logic Model Used in Evaluation of the Children At Risk Program



These intermediate outcomes, measured at the end of program participation, are hypothesized to be requisite steps towards the desired longer-term outcomes-prevention of drug use, drug selling, delinquency, school failure and dropout, and teen parenthood.

Program outputs, not shown in this diagram, include indicators of performance such as the number of tutoring sessions provided, number of home visits by case managers, and number of times parents participated in program activities.

Assessing Readiness for Evaluation

Evaluability assessment is a systematic procedure for deciding whether program evaluation is justified, feasible, and likely to provide useful information. Questions to be considered in an evaluability assessment include: (3)

Is the program's logic model plausible given the resources available and guidance from the relevant literature? If program goals are unrealistic or the intervention strategies not well grounded in theory and/or prior evidence, then evaluation is not a good investment.

What kinds of data will be needed, from what number of subjects, and what data are likely to be already available? Evaluations should be designed to maximize the use of available data, as long as these are valid indicators of important concepts and are reliable. Available data may, for example, include government statistics, individual and summary agency records and statistics, and information collected by researchers for other studies. If there are crucial data needs not met with existing data, resources must be available to collect the requisite new data.

Are adequate resources and assets available-money, time, expertise, and community and government support? Are there any factors that limit or constrain access to these resources?

Can the evaluation be achieved in a time frame that will permit the findings to be useful in making program and policy decisions by federal, state, and local officials?

To what extent does evaluation information already exist somewhere on the same or a closely related intervention? The answer to this question can have important implications for action. Any successful previous attempts may yield promising models for replication. Lessons learned from previous unsuccessful attempts may inform the current effort. If sufficient evidence already exists from previous efforts, the value of a new evaluation may be marginal.

To what extent are the findings from an evaluation likely to be generalizable to other communities, and therefore useful in assessing whether the program should be expanded to other settings or areas? Are there unique characteristics of the projects to be evaluated that might not apply to most other projects? Program characteristics that are not generalizable reduce the value of any findings.

Selecting an Evaluation Design

Selection of the evaluation design follows the systematic consideration of these questions. As noted, there are four major types of evaluation: impact, performance monitoring, process, and cost. We discuss each in turn.

Impact Evaluation Designs

Three possible designs are possible for impact evaluations: experimental, quasi-experimental, and non-experimental. They all share the strategy of comparing program outcomes with some measure of what would have happened without the program. Experimental designs are the most powerful and produce the strongest evidence. These are not always possible, however, in which case one of the two other alternatives must be chosen. (A later section discusses how to make the choice.)

EXPERIMENTAL DESIGNS

Key elements. Experimental designs are considered the "gold standard" in impact evaluation. Experiments require that individuals or groups, such as classrooms or schools, be assigned at random (by the flip of a coin or equivalent randomizing procedure) to one or more groups prior to the start of services. The "treatment" group or groups will be designated to receive particular services designed to achieve clearly specified outcomes. If multiple treatment groups are designated, the outcomes for the treatment groups may be compared to one another to estimate the relative impact of the different services or the impact relative to a control group. A "control" group receives no services. The treatment group outcomes are compared to control group outcomes to estimate impact. Because chance alone determines who receives the program services, the groups can be assumed to be similar on all characteristics that might affect the outcome measures except the program. Any differences between treatment and control groups, therefore, can be attributed with confidence to the impacts of the program.

Design Variations. One design variation is based on a random selection of time periods during which services are provided. For example, new services may be offered on randomly chosen weeks or days. A version of this approach is to use "week on/week off" assignment procedures. Although not truly random, this approach closely approximates random assignment if client characteristics do not vary systematically from week to week. It has the major advantage that program staff often find it easier to implement than making decisions on program entry by the flip of a coin on a case-by-case basis. A second design variation is a staggered start approach -in which some members of the target group are randomly selected to receive services with the understanding that the remainder will receive services at a later time (in the case of a school or classroom, the next semester or month). One disadvantage of the staggered start design is that the observations of outcomes are limited to the period between the time the first group completes the program and the second group begins. As a result, it is generally restricted to assessing gains made during participation in relatively short-term programs.

Limitations/Considerations. Although experiments are the preferred design for an impact evaluation on scientific grounds, random assignment evaluations are not always the ideal choice in real-life settings. Some interventions are inherently impossible to study through randomized experiments. Youth curfews, for example, cannot be enforced against a randomly selected subset of children in a community. And "week on/week off" enforcement is likely to breed contempt for both the law and enforcement.

A second consideration is whether random assignment is ethical and acceptable to the community. Public opinion may resist treating similar children differently on the basis of a coin flip or may view random assignment as exploiting vulnerable populations and powerless people. Carefully designed procedures for randomization may be able to overcome such resistance. One strategy is random selection of these to receive services from a list of those who meet eligibility requirements when resources are not available to serve everyone who is eligible. This form of drawing lots is close enough to "first come, first served" to be accepted as fair in many situations. Providing services for some clients at a later time (the next month or semester as described above) may satisfy community concerns about fairness and be consistent with available staff and resources. Sometimes, random assignment can involve *relaxing* a requirement instead of adding one, which makes randomization less controversial.

Great care needs to be taken to ensure that the control group is not denied essential services they would otherwise have, that the benefits to participants and the community are carefully explained, and that program staff and participants understand and support the research. Many funders require a formal review of the research design by a panel trained in guidelines developed to protect research participants. Even when such review is not required, explicit consideration of this issue is essential.

A third important issue is whether the results that are likely to be obtained justify the investment. Experiments typically require high levels of resources--money, time, expertise, and support from program staff, government agencies, funders and the community. Evaluation planners have to ask themselves whether the answers to the list of evaluation questions--and the decisions on program continuation, expansion, or modification that will be made on the basis of the findings--could be based on less costly, less definitive, but still acceptable evaluation strategies.

Practical Issues. Experimental designs run the most risk of being contaminated because of deliberate or accidental mistakes made in the field. To minimize this danger, there must be close collaboration between the evaluation team and the program staff in identifying objectives, setting schedules, dividing responsibilities for record-keeping and data collection, making decisions regarding client contact, and sharing information on progress and problems. Active support of the key program administrators, ongoing staff training and communication via meetings, conference calls, or e-mail are essential.

Failure to adhere to the plan for random assignment is a common problem. Staff are often intensely committed to their clients and will want to base program entry decisions on their perceptions of who needs, or will benefit from, the program. To prevent this pitfall, procedures should be set up so that the evaluator, not program staff, is in charge of the allocation to treatment or control group. Statistical adjustments in the analysis may be needed if there are

operational failures to maintain the randomization process(4). And even these may be inadequate to remove the biases thus introduced.

Another potential problem area is noncomparable information for treatment and control group members. Program staff can readily collect data and provide contact information for treatment group members because they have continuing contacts with clients, other agencies, and the community. Collecting comparable data and contact information on control group members can be difficult. If the experiment loses track altogether of more control than treatment group members, the evaluation data will not only be incomplete, it will provide distorted and therefore misleading information on what impacts the program has. The best way to avoid bias from this problem (called differential attrition) is to plan tracking procedures and data collection at the start of the evaluation, gathering information from the control group members on how they can be located, and developing agreements with other community agencies, preferably in writing, for assistance in data collection and sample member tracking. These agreements are helpful in maintaining sample continuity in the face of staff turnover at the agencies involved.

If the program services and content change over time, it may be difficult to determine what level or type of services produced the outcomes. The best strategy is to identify key changes in the program and the timing of changes as part of a process evaluation and use this information to define "types of program" variations in the program experience of different participants for the impact analysis. Other potential problems may be solvable through the use of special statistical techniques. Such problems include insufficient or unequal follow-up periods for treatment versus control,(5) and the risk of events (e.g., failure in school, incarceration, injury, moving) that are more likely to remove some types of members from a sample than others before the end of the planned follow-up period.(6)

Example. The evaluation of Project Alert, an eight-week junior high school curriculum for teaching seventh grade students to avoid drug use, used an experimental design(7). Thirty California and Oregon schools were randomly assigned to three groups: 1) students instructed by adult health educators, 2) students instructed by older teenagers, and 3) a no-treatment control group, although four of the non-treatment schools provided other drug prevention instructional programs. To increase the generalizability of the findings, the schools were drawn from eight urban, suburban, and rural communities and nearly a third of the schools had minority populations of 50 percent or higher. To increase the pre-assignment similarities of the three experimental groups and strengthen the statistical power of the analysis (given the relatively small sample of schools), each experimental group was included in at least one school in each community, and the schools included in the experiment were matched to the extent possible to reduce differences among groups in such characteristics as test scores, language spoken at home, drug use among 8th graders, and ethnic and income composition. These procedures produced substantial pre-experimental similarities in factors related to drug use among the experimental groups. Since schools but not students were randomly assigned, statistical adjustments were used to correct for the clustering of students within schools. Students completed questionnaires about their drug use seven times between grades 7 and 12; those who transferred to other schools or districts completed mail and telephone interviews to minimize sample attrition. Outcome measures included cognitive risk factors associated with drug use: beliefs about consequences of use, norms regarding drug use, peer resistance, self-efficacy, and expected future drug use.

Experimental evaluations are costly. The Children At Risk evaluation, for example, cost \$1.5 million. But the rigorous design permitted strong conclusions about the long-term effectiveness of drug prevention education during early adolescence and demonstrated that results are not restricted to middle class communities, but can be used in schools with high proportions of lower income and minority students.

QUASI-EXPERIMENTAL DESIGNS

Key Elements. Like experiments, quasi-experimental evaluations compare outcomes from program participants to outcomes for comparison groups that do not receive program services. The critical difference is that the decision on who receives the program is not random. Comparison groups are made up of members of the target population as similar as possible to program participants on factors that could affect the selected outcomes to be observed. Multivariate statistical techniques are then used to control for remaining differences between the groups.

Usually, evaluators use existing population groups for comparison—those who live in a similar area, or are enrolled in the same school in a different classroom, or attended the same school with the same teacher in the previous year. In some situations, staff (or schools or communities) are willing or trained to try the new "treatment" while others are not, but the same rules for service eligibility are used by all.

Design Variations. The primary variation is to construct a comparison group by matching individuals to individuals in the treatment group on a selected set of characteristics. This process for selecting a comparison group is methodologically less defensible(8). The threats to validity are twofold. 1) Matches based on similarities at a single point in time do not always result in groups of individuals who are comparable over time. Thus, the groups may become increasingly different over time independent of the program. 2) Differences in variables not used in the matching may have a substantial effect independently of the program being evaluated.

Quasi-experimental designs vary in the number and timing of the collection of data on program outcome measures. The selection of the number and timing of measurements is based on an assessment of the potential threats posed by competing hypotheses that cannot be ruled out by the comparison methodology. In many situations, the strongest designs are those that collect pre-program measures of outcomes and risk factors and use these in the analysis to focus on within-individual changes that occur during the program period. These variables are also used to identify groups of participants who benefit most from the services. One design variation involves additional measurement points (in addition to simple before and after) to measure trends more precisely. Another variation is useful when pre-program data collection (such as administering a test on knowledge or attitudes) might "teach" youth about the questions to be asked after the program to measure change, and thus distort the measurement of program impact. This variation involves limiting data collection to the end of the program period for some groups, allowing their post-program answers to be compared with the post-program answers of those who also participated in the pre-program testing.

Considerations/Limitations. Use of non-equivalent control group designs requires careful attention to procedures that rule out competing hypotheses regarding what caused any observed differences on the outcomes of interest. In evaluations of programs for vulnerable children and youth, three threats to validity stand out.(9)

The first is the threat of "maturation"--the possibility that age-related processes will contribute to outcomes independently of the program intervention. Among youth, certain outcomes, positive and negative, are strongly tied to age--outcomes such as drug use, delinquency, and early parenthood. It is therefore necessary to be sure that the comparison group is made up of youth at the same developmental stage.

A second threat is that of "history"--the risk that unrelated events may affect outcomes. For example, the rapid spread of crack use among women childbearing age in the United States in the late 1980s greatly increased rates of drug-exposed infants. Thus, a comparison group for an evaluation of a prenatal health care program would need to be drawn from the same years and communities to "control" for the spread of crack. Otherwise, the upward trend in negative outcomes due to crack could obscure the prevention benefits of the program. Similarly, designs need to consider controls for geographic variation in events external to the program. For example, gang crackdowns in some neighborhoods and not others could influence assessments of the impact of a school-based delinquency or drug prevention program. If the crackdown occurred in the "treatment" neighborhood, the program effects might be over-estimated; if it occurred in the comparison neighborhood, program effects might be under-estimated.

A third threat to validity is the process of "selection" -the factors that determine who receives services. Some of these factors are readily identified and can be used as control variables in statistical models, such as living in a specific school district or meeting program eligibility criteria. However, it is unlikely that all factors will be correctly identified and adequately measured. For example, program participants may receive services because they are more motivated, skillful, or socially well-connected than nonparticipants. Such differences are not easy to measure during a program evaluation.

Practical Problems. Building defenses or "controls" for threats to validity into evaluation designs through the selection of comparison groups and the timing of outcome observations is a challenge. Controls for maturation, history, and selection may involve, respectively, selecting a sample that includes multiple age cohorts, collecting data in similar or nearby localities that lack the program,(10) or applying a statistical model that controls for foreseeable biases in selecting program participants.(11) Even when the comparison group is carefully selected, the researcher cannot be sure that all relevant group differences have been identified and measured accurately. Statistical methods can adjust for such problems and increase the precision with which program effects can be estimated,"(12) but they do not fully compensate for the non-random design. Findings need to be interpreted extremely cautiously and untested alternative hypotheses carefully considered.

As in experimental evaluation, plans for quasi-experimental evaluations need to pay close attention to the problem of collecting comparable information on control group members and developing procedures for tracking them. However, the need for close collaboration with

program staff is reduced, since the staff are generally neither involved in selecting participants nor in contact with comparison group members.

Example. The evaluation of the Teen Age Parenting Program (TAPP) for adolescents divided teen mothers into three groups designed to be similar in age and other characteristics.(13) Each group was evenly divided among black, Hispanic, and white participants. One group attended an alternative school with child development and parenting classes and a nursery school featuring a parenting-child development curriculum. Another group attended an alternative school without a nursery school. The remaining group received no special services for teenage parents. Services began during pregnancy. Assessments of educational progress, fertility, knowledge, and child development two to four years later were based on interviews and school records. Mothers in the alternative school with the nursery program had completed more schooling and were more likely to still be enrolled in school than the other mothers. Mothers in both alternative schools had more knowledge about parenting and reproduction and more positive attitudes about parenting than those without special services. But there were no significant differences in the groups on child development outcome measures. How to interpret, this seeming inconsistency is complicated, because the evaluation design did not have pre-program measures of individual differences and assignment was not random. The education and knowledge differences across the three groups may have been there from the beginning, rather than being attributable to the special services.

NON-EXPERIMENTAL IMPACT EVALUATIONS

Key Elements. Non-experimental impact evaluations examine changes in levels of risk or outcomes among program participants, or groups including program participants, but do not include comparison groups of other individuals or groups not exposed to the program.

Design Variations. The four primary types of non-experimental designs include: 1) before and after comparisons of program participants; 2) time series designs based on repeated measures of outcomes before and after the program for groups that include program participants; 3) panel studies based on repeated measurement of outcomes on the same group of participants; and 4) post-program comparisons among groups of participants.

The first two designs are based on analysis of aggregate data. In **before and after comparisons**, outcomes for groups of participants (program groups that enter the program at a specific time and progress through it over the same time frame) are measured before and after an intervention and an assessment of impact inferred from the differences. This simple design is often used to assess whether knowledge, attitudes, or behavior of the group changed after exposure to a classroom curriculum or job training program. **Time series designs** are an extension of the before and after design that uses multiple measures of the outcome variables before an intervention begins and continues to take multiple measures after intervention is in place. If a change in the trend (direction or level) in the outcome occurs at, or shortly after the time of the intervention, the significance of the observed change is tested statistically. Time series measures may be based on larger groups or units that include but are not restricted to program participants. For example, crime rates for neighborhoods in which most or all youth participate in a delinquency prevention program might be used to assess reductions in illegal activity. Evaluation

of a series of dropout prevention activities offered across the school year could examine the percentages of entering classes that graduate over a period of years. Time series designs should be considered when it is difficult to identify who receives program services or when the evaluation budget does not support collection of detailed data from program participants. Although new statistical techniques have strengthened the statistical power of these designs,(14)" it is still difficult to rule out the potential impact of non-program events using this approach.

The next two designs examine data at the individual level. **Cross-sectional comparisons** are based on surveys of groups of participants conducted after program completion. This design can be used to estimate correlations between outcomes and differences in the duration, type, and intensity of services received, yielding conclusions about plausible links between outcomes and services but no definitive conclusions about what caused what. **Panel designs** use repeated measures of the outcome variables for each individual. In this design, outcomes are measured for the same group of program participants, often starting at the time they enter the program and continuing at intervals over time. For example, the evaluation of Health Planning and Promotion: Life Planning Education used pre-post data from participants to measure gains in understanding the best combinations of contraceptive methods and the consequences of early childbearing.(15) This design allows the characteristics of individual participants to be used in the analysis to identify different patterns of change associated with individual characteristics of participants and control for other events to which they were exposed.

Considerations/Limitations. Several limitations to non-experimental designs should be noted. First, the cross-sectional and panel designs provide only a segment of "dose-response curve," that is, only estimates of the *differences* in impact related to differences in the services received. These designs cannot estimate the full impact of the program compared to no service at all, unless estimates can be based on other information on the risks of the target population. Second, the designs that track participants over time (before and after, panel, and time series) cannot control for the effects of developmental changes that would have occurred without services, or for the effects of other events outside the program's influence. Third, the extent to which the results can be assumed to apply to other groups or other settings is limited, because this design provides no information for assessing the extent to which participants were selected into the program on the basis of factors which themselves influence outcomes.

Practical Issues. Non-experimental designs have considerable practical advantages because they are relatively easy and inexpensive to conduct. Individual data for cross-sectional or panel analysis are often collected routinely by the program at the end (and sometimes beginning) of program participation. When relying on program records, the evaluator needs to review the available data against the logic model to be sure that adequate information on key variables is already included, or to begin collecting additional data items if needed. When individual program records are not available, aggregate statistics may be obtained from the program or from other community agencies with information on the outcomes among groups of participants. For example, crime rates, average promotion rates, and rates of births to teen mothers can be collected from existing records. The primary problem encountered in using such statistics for assessing impacts is that they may not be available for the specific population or geographic area targeted by the program. Often these routinely collected statistics are based on the general population or geographic areas served by the agency (e.g., the police precinct or the

clinic catchment area). The rates of negative outcomes for the entire set of cases included may well differ from rates for the targeted group of vulnerable children and youth; this risk is greater for larger rather than smaller statistical areas.

A more expensive form of data collection for non-experimental evaluations is a survey of participants some time after the end of the program. These surveys can provide much needed information on longer-term outcomes such as rates of employment or earnings or high school graduation. As in any survey research, the quality of the results is determined by response rate rather than overall sample size, and by careful attention to the validity and reliability of the questionnaire items.

Example. The Youth Training Scheme (YTS) in Great Britain provides, through local agents, two years of vocational and on-the-job training for out-of-school and unemployed youth ages 16 and 17. The local agents are businesses or community organizations that receive government funds to design a training program, recruit and supervise youth, and provide at least 13 weeks of on-the-job training per year. Non-experimental evaluation of YTS was based on a follow-up survey of 63,000 former participants.⁽¹⁶⁾ In addition to monitoring client satisfaction and job related outcomes, the survey was used in non-experimental comparisons of differences in outcomes related to differences among participants: job market outcomes were compared for graduates versus program dropouts and across youth who entered the program with different levels of motivation and past school achievement. Results indicate that program graduates had better labor market outcomes than those who did not complete the program. Similarly, earning qualifications in the program (an interim outcome measure) was positively correlated with later labor market success (the longer term outcome). Non-experimental comparisons were also used to identify differences in outcomes related to characteristics of the participants or the training experience. The field of employment and type of local agent providing the training were significant predictors of labor market outcomes. Similarly, labor market outcomes were better for youth who began the program with higher levels of motivation and past school achievement. These findings are suggestive but not definitive. Because of the non-experimental design, participating youth might have been more likely to become employed than other youth even in the absence of the program.

CHOOSING AMONG THE IMPACT DESIGNS

Choice of an impact evaluation design begins by identifying the design that both offers the strongest capacity for isolating the independent causal effects of the program and is feasible given the structure of program. The "decision tree" shown in Exhibit B illustrates a process for identifying which alternatives are feasible.

If the program will be provided to a limited number of youth who can be identified in advance, and randomly selected for participation, then an experimental design should be considered. If the program will be provided to a limited number of youth, but the decision about who receives services is determined by organizational or geographic considerations (or other nonrandom selection rules), then quasi-experimental design variations should be considered.

The most difficult design challenges occur when the program is intended to serve all members of the target population. If the new program is implemented fully and rapidly, no youth will be available for a comparison group. Often, however, new full-coverage programs—for example, new health services—are intended for an entire population but not implemented in every community in the country, and certainly not at the same time. If some communities or groups are not included in the initial implementation, it may be possible to select as comparison sites communities that have not implemented the program and use a quasi-experimental design. This may not solve the problem of comparability sufficiently to allow such a design, however, if the communities where it was implemented have characteristics that are systematically different from those where it was not.

When non-experimental designs are necessary, the following can help guide the choice of design. If a program is implemented at different levels across sites but uniformly *within* sites, a cross-sectional design is suitable. If a target population is exposed to different levels of the program within a community, a panel study design is better—to follow a sample of individuals, and record both outcomes and the amount of the program or intervention each individual received and when it occurred. If defining who is served by the program is difficult or the program is uniformly applied in all communities, then a time-series design is appropriate. Before-and-after designs without control groups are often used, but are subject to a number of threats to validity, including maturation and secular changes (discussed above).

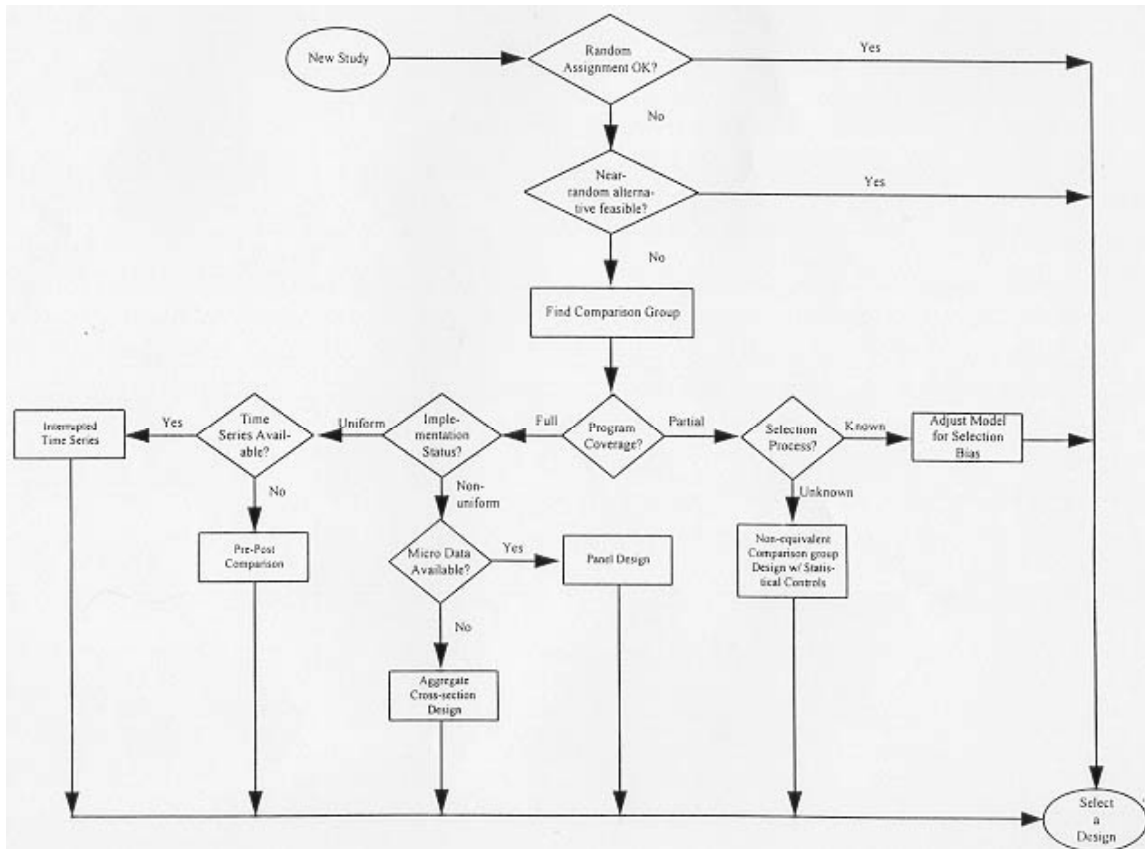
Performance Monitoring

Key Elements. Performance monitoring is used to provide information on: 1) key aspects of how a system or program is operating; 2) whether, and to what extent, pre-specified program objectives are being attained (e.g., numbers of youth served compared to target goals, reductions in school dropouts compared to target goals); and 3) identification of failures to produce program outputs, for use in managing or redesigning program operations. Performance indicators can also be developed to 4) monitor service quality by collecting data on the satisfaction of those served, and 5) report on program efficiency, effectiveness, and productivity by assessing the relationship between the resources used (program inputs) and the output and outcome indicators.

If conducted frequently enough and in a timely way, performance monitoring can provide managers with regular feedback that will allow them to identify problems, take timely action, and subsequently assess whether their actions have led to the improvements sought. Performance measures can also stimulate communication about program goals, progress, obstacles, and results among program staff and managers, the public, and other stakeholders. They focus attention on the specific outcomes desired and better ways to achieve them, and can promote credibility by highlighting the accomplishments and value of the program.

Performance monitoring involves identification and collection of specific data on program outputs, outcomes, and accomplishments. Although they may measure subjective factors such

Process for Selecting Impact Evaluation Designs



as client satisfaction, the data are numeric, consisting of frequency counts, statistical averages, ratios, or percentages. Output measures reflect *internal* activities: the amount of work done within the program or organization. Outcome measures (immediate and longer term) reflect progress towards program goals. Often the same measurements (e.g., number/percent of youth who stopped or reduced substance abuse) may be used for performance monitoring and impact evaluation. However, unlike impact evaluation, performance monitoring does not make any rigorous effort to determine whether these were caused by program efforts or by other external events.

Design Variations. When programs are operating in a number of communities, the sites are likely to vary in mission, structure, the nature and extent of project implementation, primary clients/targets, and timeliness. They may offer somewhat different sets of services, or have identified somewhat different goals. In such situations, it is advisable to construct a "core" set of performance measures to be used by all, and to supplement these with "local" performance indicators that reflect differences. For example, some youth programs will collect detailed data on youth school performance, including grades, attendance, and disciplinary actions, while others will simply have data on promotion to the next grade or whether the youth is still enrolled

or has dropped out. A multi-school performance monitoring system might require data on promotion and enrollment for all schools, and specify more detailed or specialized indicators on attendance or disciplinary actions for one or a subset of schools to use in their own performance monitoring.

Considerations/Limitations. In selecting performance indicators, evaluators and service providers need to consider:

The relevance of potential measures to the mission/objective of the local program or national initiative. Do process indicators reflect program strategies/activities identified in mission statements? Do outcome indicators cover objectives identified in mission statements? Do indicators capture the priorities at the community level?

The comprehensiveness of the set of measures. Does the set of performance measures cover inputs, outputs, and service quality as well as outcomes and include relevant items of customer feedback?

The program's control over the factor being measured. Does the program have influence/control over the outputs or outcomes measured by the indicator? If the program has only limited influence over the outputs or outcomes being measured, the indicator may not fairly reflect program performance.

The validity of the measure. Do the proposed indicators reflect the range of outcomes the program hopes to affect? Are the data free from obvious reporting bias?

The reliability and accuracy of the measure. Can indicators be operationally defined in a straightforward manner so that supporting data can be collected consistently over time, across data gatherers, and across communities? Do existing data sources meet these criteria?

The feasibility of collecting the data. How much effort and money is required to generate each measure? Should a particularly costly measure be retained because it is perceived as critically important?

Practical Issues. The set of performance indicators should be simple, limited to a few key indicators of priority outcomes. Too many indicators burden the data collection and analysis and make it less likely that managers will understand and use reported information. At the same time, the set of indicators should be constructed to reflect the informational needs of stakeholders at all levels-community members, agency directors, and national funders.

Regular measurement, ideally quarterly, is important so that the system provides the information in time to make shifts in program operations and to capture changes over time. However, pressures for timely reporting should not be allowed to sacrifice data quality. For the performance monitoring to take place in a reliable and timely way, the evaluation should include adequate support and plans for training and technical assistance for data collection. Routine quality control procedures should be established to check on data entry accuracy and missing

information. At the point of analysis, procedures for verifying trends should be in place, particularly if the results are unexpected.

The costs of performance monitoring are modest relative to impact evaluations, but still vary widely depending on the data used. Most performance indicator data come from records maintained by service providers. The added expense involves regularly collecting and analyzing these records, as well as preparing and disseminating reports to those concerned. This is typically a part-time work assignment for a supervisor within the agency. The expense will be greater if client satisfaction surveys are used to measure outcomes. An outside survey organization may be required for a large-scale survey of past clients; alternatively, a self-administered exit questionnaire can be given to clients at the end of services. In either case, the assistance of professional researchers is needed in preparing data sets, analyses, and reports.

Example. The Asociacion Salud con Prevencion (ASCP) in Colombia, South America, a non-governmental organization which provides primary prevention services which promote adolescent reproductive health, monitors outputs with data on the number of professionals trained, the number of youth given educational services, the number of workshops held, the number of condoms distributed, and the number of medical and counseling sessions provided. The results demonstrate that the program is providing promised services, but does not give an indication of the impact in terms of either immediate outcomes such as use of birth control or longer-term outcomes (which include reduced risk of out-of-wedlock births or early childbearing).

Process Analysis

Key Element. The key element in process analysis is a systematic, focused plan for collecting data to: (1) determine whatever the program model is being implemented as specified and, if not, how operations differ from those initially planned; (2) identify unintended consequences and unanticipated outcomes; and (3) understand the program from the perspectives of staff, participants, and the community.

Design Variations. The systematic procedures used to collect data for process evaluation often include case studies, focus groups, and ethnography.

Case studies involve the detailed analysis of selected program sites or clients to determine how the program is operating, what barriers to program implementation have been encountered, what strategies are the most successful, and what resources and skills are necessary. The answers to these questions are useful in providing guidance to policymakers and program planners interested in identifying key program elements and in generating hypotheses about program impact that can be tested in impact analyses. Case studies are sometimes used to test competing hypotheses about differences in the impact of services. This strategy is used to assess which approach is most successful in attaining goals shared by all when competing models have emerged in different locations. This requires purposely selecting sites to represent variations in

elements or types of programs, careful analysis of potential causal models, and the collection of qualitative data to elaborate the causal links at each site.

Clients or sites chosen for case studies should represent wide variation in settings, program models, and clients. Identification of sample members within sites, interview topics, and key data elements begins with the logic model as a guide. In a case study, qualitative data, collected using semi-structured interviews and observations of program operations, are often supplemented and verified by quantitative data on program operations and performance collected from records and reports.

Case studies may use several different approaches for collecting qualitative data for program evaluation. The most frequently used are semi-structured interviews, focus groups, and researcher observations while on-site. Semi-structured interviews allow for the discovery of unanticipated factors associated with program interpretation and outcomes. Protocols for semi-structured interviews contain specific questions about particular issues or program practices. The "semi" aspect of these discussion guides refers to the fact that a respondent may give as long, detailed, and complex a response as he or she desires to the question-whatever conveys the full reality of the program's experience with the issue at hand. If some issues have typical categories associated with them, the protocols will usually contain probes to make sure the researcher learns about each category of interest.

In case studies, observations at program sites provide an important method of validating information from interviews. In this case, the observations will often be guided by structured or semi-structured protocols designed to ensure that key items reported in interviews are verified and that consistent procedures for rating program performance are used across time and across sites.

Focus groups seek to understand attitudes through a series of group discussions guided by one researcher acting as a facilitator, with another researcher present to take detailed notes. Five or six general questions are selected to guide open-ended discussions lasting about an hour and a half. The goals of the discussions may vary from achieving group consensus to emphasizing points of divergence among participants. Discussions are tape-recorded, but the primary record is the detailed notes taken by the researcher who acts as recorder. Less detailed notes may also be taken publicly, on a flipchart for all to see, to try to achieve consensus or give group members the chance to add anything they think is important. Soon after a particular focus group, the recorder and facilitator summarize in writing the main points that emerged in response to each of the general questions. When all focus groups are completed, the researchers develop a combined summary, noting group differences and suggesting hypotheses about those differences.

Ethnography relies almost exclusively on observation and unstructured interviews to study:

- * Organizational and programmatic processes occurring at a program site;
- * The community context in which the program is taking place;

- * The relationship between program activities and other activities in the community;
- * Causal processes as the participants view them; and
- * Modes of decision-making.

Ethnography does not begin with the logic model. Its intent is to understand the program from the perspective of staff, participants, and others in the community. Ethnographers observe program operations as unobtrusively as possible, sometimes in the role of participant observer, and keep detailed field notes that are transcribed and coded to identify emerging themes and trends. The critical research goal is to provide data on the subjective experience of those in the program situation and to use this information to understand if the program goals are being achieved and, if so, how.

Ethnography uses procedures that are deliberately flexible. As a result, ethnography is helpful in gathering information on unintended consequences and unanticipated outcomes. These unexpected observations may lead to an entirely new concept of program delivery. In a recent project examining service integration programs for at-risk youth, observations helped clarify that service integration needed to go beyond formal links and on-paper agreements, and provided insights into how informal processes bonded services together in their efforts to make a difference for high-risk youth in the community.⁽¹⁷⁾ Observations from ethnographic studies are perhaps the hardest type of qualitative information to analyze, since they generate volumes of information, much of which may not be directly related to evaluation goals and may not be comparable across sites.

Practical Issues. Collecting qualitative data requires skilled researchers who are experienced with the techniques being used. To analyze these data, careful notes must be taken to ensure that responses are correctly recorded and to aid in interpreting them. In methods based on interviews, interviewers must be trained to understand the intent of each question, the possible variety of answers that respondents might give, and ways to probe to ensure that full information about the issues under investigation is obtained.

Analysis of qualitative data requires an in-depth understanding of programs, respondents and responses, and especially the context in which they are evaluated. Ultimately, the analyst makes judgments regarding the relative importance or significance of various responses. This requires an unbiased assessment of whether responses support or refute hypotheses about the way the program works and the effects it has.

One way to handle qualitative data is to treat one's interview and observational notes as text, and to conduct a textual analysis using specialized computer software that can search for the presence of specific themes or content. Qualitative software is available to facilitate the location and retrieval of information from massive textual files. This kind of software is expensive to use because huge amounts of text must be entered into a computer. Further, either the exact words one wants to search for must appear in the text, or the text marked for the presence of any theme or topic that the researcher wants to retrieve. Often researchers can achieve equal or better results with carefully constructed interview or data collection guides or structured focus groups, and systematically recording of responses or coding of data encountered in the field.

Example. Case studies of two pilot projects were used for the evaluation of mentoring in the juvenile justice system conducted by Public/Private Ventures. The program was designed to match 100 mentors to at-risk youth. Mentors were trained to meet with youth one-on-one before and after the youth's release from juvenile detention facilities, with the goal of establishing an attachment to an adult role model. Data were collected from mentor logs, program records, court records, structured interviews with mentors and youth before and after program participation, staff interviews, focus groups with mentors, youth and service agency staff, and in-depth interviews with mentor-youth pairs. The qualitative analysis examined the characteristics of successful matches, issues in program implementation, the style and content of mentoring interactions, and program staffing. Although it does not offer evidence on outcomes, the evaluation provides extremely useful information on the process of implementing a mentoring program and guidance for program development and replication.

Cost Studies

Key Elements. Cost studies are used to assess investments in programs by collecting information on: 1) direct program expenditures; 2) the costs of staff and resources provided by other agencies or diverted from other uses; 3) costs for purchased services; and 4) the value of donated time and materials. Costs for the first two items usually include expenditures for staff salaries; fringe benefits; special training costs (if any); travel; facilities; and supplies and equipment that have to be purchased. The value of donated resources, which can be substantial, generally has to be estimated and requires careful documentation of the donation. Cost analyses indicate that donations are a major cost item in many youth programs. For example, the Cities in Schools (18) evaluation indicated that donations are between 74 percent and 90 percent of the total direct program costs, and that the wide variation among cities in the types of donations received made the inclusion of these costs essential to an understanding of the resources required to sustain program operation.

The typical approach to cost studies is to calculate total program costs and then an *average* cost per client, calculated by dividing the total by either the total number of clients served, or the total number of clients who meet some standardized definition of success. This type of cost calculation can be linked to results of an experimental or quasi-experimental impact evaluation to estimate costs per successful client. It can also be used with performance indicators to assess the cost or cost-efficiency of achieving program goals.

A second approach to cost estimation calculates the cost per unit of service. For example, the cost per hour of classroom instruction or the cost per hour of counseling. This type of cost calculation is then used in impact evaluations (including non-experimental evaluations) to look at the costs of different outcomes. This type of cost analysis is difficult in multi-faceted, comprehensive programs in which the level and type of service are highly variable and may involve a number of service providers. It is also difficult in programs in which defining exposure to services is difficult. Where possible, it is preferable to distinguish between fixed costs (e.g., rent or the director's salary) and variable costs (e.g., the costs of special events or the hourly

costs of the recreation director). The variable costs can then be used to estimate the *marginal* cost of adding additional clients to the number receiving a specific unit of service.

Design Variations. Cost studies can be undertaken to describe the program costs and link these to the level of outcomes achieved. In this application, the costs are compared to the level and type of outcomes documented in performance monitoring outcomes. Decisions on whether the outcomes justify the costs are based on opinions about the value of the outcomes (not monetized) and the likelihood that the outcomes are attributable to the program.

Cost-effectiveness analysis is used to compare the costs of different approaches to providing some standard level of service or desired level of outcome. This approach is most useful when multiple programs are using different models to provide a service. The requirements are that the characteristics of target populations served, the program goals, and the output or outcome measures be identical. For example, cost-effectiveness studies could compare the relative effectiveness of residential and nonresidential treatment for drug-abusing youth, provided that the youth served were similar in age and drug use problems, and that the same measures of treatment success were used.

Cost-benefit studies provide estimates of the dollar benefits returned for each dollar spent on the program—the key question from a policy perspective, but one that is not easily answered. This type of evaluation has rigorous requirements for: 1) an estimate of program costs, either per client or per unit of service; 2) estimates of the value of the benefits; and 3) comparative data on program impact—an estimate of outcomes with and without the program. The first item should be obtainable from program financial records, supplemented as needed by estimates of the cost of donated or reallocated resources. The second can be obtained from an experimental or quasi-experimental evaluation of program impact or another strategy for estimating the difference between what happened and what would have happened without the program.

The primary barrier to conducting cost-benefit analysis of service programs designed to change behavior stems from the third item: placing dollar values on benefits. Many benefits are of intrinsic value (e.g., reductions in family dysfunction and conflict) but quantifying that value is difficult.

Monetization of *benefits to individuals* requires assumptions about three matters, all of which are frequently controversial. First, the dollar value of the benefit may depend on personal values, for example, what residents are willing to pay for a crime-free neighborhood. Second, a dollar of benefit today is worth more than a dollar benefit realized next year. Thus, the benefits need to be time discounted, but by how much is a difficult question. Third, the beneficiaries need to be identified. Societal values become important when the beneficiaries differ in standing and perceived merit. For example, a high school equivalency degree for a violent youthful offender may result in the same gains in lifetime earnings for the offender as a violence victim would realize from physical therapy for the injury. Are they to be treated the same? To circumvent such difficult questions, the analyst may conduct a sensitivity analysis to reach conclusions based on explicit assumptions of value. For example, the neighborhood crime prevention program may be deemed cost-effective if "residents are willing to pay at least \$100 per month for 10 percent

lower rates of burglary" or "if the discount rate is less than 6 percent" or "if the offender's earnings are worth 50 percent of the victim's earnings."

Beyond benefits to individuals, the total value of benefits includes the *social costs averted*. These are the savings to the public that result from avoiding negative outcomes. These values must be based on studies that estimate the social costs of negative outcomes such as the costs of crime or drug abuse.(19)" These estimates are difficult to derive and are often based on tenuous assumptions. To compensate for problems in the reliability of estimates, cost-benefit calculations normally use a range of benefits to place an upper and a lower bound on the probable returns to investments in the program. A more significant problem is that monetary values based on public costs for the negative outcomes among the general population may be poor estimates of the value of benefits among the program's *target* population. For example, national estimates of the costs of drug abuse may not apply to reductions in amphetamine abuse among low-income adolescents in a single city. This problem needs to be acknowledged and value estimates revised to the extent possible to reflect savings for the program's participants. Other public benefits reflecting gains, not costs averted, are widely acknowledged, but rarely find their way into cost-benefit studies because there is no public consensus on their importance. Examples include improvements in the quality of life or the environment.

Considerations and Limitations. Documentation of gains to prevention programs is exceptionally difficult and requires estimating negative outcomes that did not occur. As described above, the most robust estimates of program impacts of this kind are based on experimental evaluations or quasi-experimental evaluations, which are difficult and expensive to conduct. When the program has total population coverage, it is possible to interpret differences between the observed trend and predicted trend in an outcome indicator over time to program impacts and estimate the monetary value of the benefits. This strategy was used to estimate the value of drug prevention efforts in the United States. National survey estimates of drug use in 1979 were used to estimate expected drug prevalence during the 1980s and early 1990s; the differences between these estimates and drug use prevalence rates based on national surveys during these years were attributed to federal investments in drug prevention programs.(20)

Practical Issues. Developing a conceptual framework that reflects all the issues in cost-benefit valuation, and then devoting the resources necessary for estimating the range of benefits, can require as much research time and expertise as determining whether the program had any impact. However, research dollars are always limited and evaluating program impact is usually the top priority, since valuing benefits is irrelevant if there is no program impact. A number of studies of the value of preventing negative outcomes among children and youth have been initiated recently. These can be expected to give program evaluators substantial help in estimating the value of reductions in youth problems for use in cost-benefit studies in the future.

Example. An evaluation of 13 delinquency prevention programs in Los Angeles County estimated cost effectiveness as a function of the delinquency risk of the population of youth served, costs, and success rate. This study compared cost to benefit ratios of alternative programs designed with a common goal and outcome measure-preventing subsequent arrest. Because the risks of delinquency varied among the youth served by different programs, estimates of the risk of delinquency was derived from existing research and used to classify the youth served by the

program into four risk categories. Program costs were estimated by taking the total budgets from all sources divided by the number of clients. Costs of public expenditures for delinquency (costs to the community and justice system) were estimated from the proportion of the justice system budget (from the County budget) devoted to juvenile cases, divided by the number of juvenile cases at various stages of processing (from annual reports of the Los Angeles Probation Department, the California Youth Authority, and the U.S. Department of Justice).

The public costs averted were calculated by dividing the budget by the number of arrests of youth following program participation and calculating the savings as the difference between the two. The benefits of reductions in expected future arrests were estimated based on the probability of subsequent arrests reported in studies of criminal careers times the estimated public savings per arrest averted. Savings to victims were based on estimates of the costs of damage and loss for each type of juvenile offense from earlier research, adjusted for inflation. These costs per offense were applied to the expected lifetime arrests in the absence of the program and benefits were estimated as the difference between these costs and the absence of costs associated with no further arrests or victimization (estimating that for each arrest, there are four to five offenses that do not result in arrest). Thus, estimated program benefits were the sum of the public costs averted and the savings to victims.

The results were used to estimate the cost differential (costs divided by the value of benefits) to programs with different rates of success (measures as arrests prevented), controlling for the risk of offending of the juvenile population served. The findings were used to estimate the success rate required to show a positive rate of return given the delinquency risk of the population served for programs with different cost differentials. This estimate can be used in monitoring the performance of a wide variety of delinquency prevention programs.

Identifying Potential Evaluation Problems

A number of challenging problems face those who would apply research methods to the evaluation of human services programs. We summarize these, based on experience in reviewing and evaluating programs for vulnerable children and youth, to guide development of realistic evaluation plans. (21)

Defining Program Participation. Programs may be open-ended, lacking both formal intake procedures and policies for determining when the program is "completed." An evaluation can only yield interpretable results if participation is explicitly defined and uniformly measured. In the case of programs for vulnerable youth, for example, counselors may be contacted for several chats, followed some weeks later by an appointment, followed by intermittent participation in some, possibly not all, services offered. Youth may stop attending and then resume. Limiting participation in the evaluation to those who attend regularly is not an appropriate solution because dropping from consideration the youth who are most difficult to engage produces biased results. Often identifying who "participated" and for how long requires multiple categories to adequately reflect the variations in type, duration, and intensity of participation among the youth served. In addition, participants should be followed from the point of first contact and all major program activity documented. Evaluators also need to decide whether others who potentially

benefit from the program-such as parents, boyfriends/girlfriends, or siblings-are defined as program participants. If so, their participation in program activities should also be tracked. If not, plans need to be made on how to count the gains made by these *indirect* program beneficiaries in evaluating program impact.

Evaluating the Relationship between Participation and Outcomes. Many programs emphasize individualized services tailored to need. In the youth services area, youth with the highest levels of risk are offered the greatest number or most intensive level of services. Obviously, assignment to treatment in this case is not random, and the multi-problem youth may never achieve the same level of positive outcomes as youth who began with fewer problems. For example, studies of the School-Based Health Centers in the U.S. show that frequent clinic users were at greater risk for alcohol and substance use, sexual activity, and poor family and peer relationships(22). Thus, comparing their outcomes to those for nonusers or those who used the clinic less frequently would be inappropriate. Similarly, comparisons between different programs must consider any differences in type and level of risk exhibited by participants. For this reason, data on the risks and needs of participants should be collected at intake for use in analysis and a pre-post design used when possible.

Defining the Unit of Analysis. Deciding on the appropriate unit of analysis can be difficult, particularly in evaluating comprehensive programs. Programs may target entire neighborhoods, classrooms, or families for change-sometimes planning activities directly for different groups, and sometimes planning carryover effects. Measurement at multiple levels is appropriate as long as each level is clearly defined. For example, crime reduction can be assessed by comparing neighborhood rates of calls for police services, household victimization rates, or youth delinquency surveys. Economic gains can be measured by changes in the area unemployment rate, average household or family income, or individual earnings. The selection should be closely linked to program goals and activities.

Evaluations of services integration programs, including most that use a case management approach, will face additional challenges in: 1) tracking the services received by participants; 2) developing common agreements among agencies on program goals and required components; 3) documenting service delivery by multiple agencies; 4) measuring effects of the service delivery system; and 5) differentiating services integration from service comprehensiveness. Each is discussed briefly below.

Tracking the services received by participants. Services integration usually involves referring participants to other agencies for needed assistance. A critical, and often difficult, problem is determining which services ' were actually received. Clients may or may not contact agencies to which they are referred, may or may not be accepted for services, and may or may not participate in services, if accepted. Documenting the chain of participation is essential to determine the extent to which services integration is being achieved, but is time consuming and often resisted by programs that see making the referral as the extent of their responsibility. Because staff turnover in service agencies is frequently high, preparing written agreements on data access and sharing is strongly recommended. In the absence of adequate agency documentation, information on service utilization can be collected in follow-up interviews with clients.

Developing common agreements among agencies on program goals and required components. The agencies collaborating in a services integration effort may differ in their vision of the program's goals, key strategies, and how youth needs will be evaluated and problems addressed. Evaluations tend to highlight these differences, which can constitute a barrier in gaining consensus on what is being evaluated. This is particularly true when multiple agencies recruit clients and/or case management services are not centralized. Time should be allocated for face-to-face meetings to get agreement on whom evaluators will count in selecting measures of program outcomes, and how service provision is expected to achieve program goals.

Documenting service delivery by multiple agencies. When many agencies coordinate and combine their resources to meet the needs of clients, one of the most difficult problems is assembling information on who received what types and amounts of service. Agencies have different methods of identifying clients. In the area of vulnerable children and youth, some use family identification numbers, others identify individual children served. Some group service records by family or child; others maintain records by contact, which introduces multiple records for single clients which then have to be checked to remove duplication. Agencies such as schools or juvenile courts can face legal or professional barriers to sharing client-based information with other agencies or evaluators. A systematic system for collecting the data needed to compile a complete picture of program participation must be developed early in the planning process and, as noted above, supported by written agreements and ongoing technical assistance and staff training in record-keeping procedures.

Measuring effects of the service delivery system. A primary goal of services integration is to change agency operations and increase effectiveness. These outcomes need to be measured at the agency, not individual, level. Evaluations of services integration need to document changes in agency procedures, increased participation in collaborative planning and service delivery, and decreases in barriers to interagency cooperation and client service associated with policies, and procedures. Referral patterns should show more diversity in planning. At the individual level, clients should report fewer unmet service needs, shorter waiting periods for service, and increased satisfaction with the response to their needs. Other evidence of integration includes increased staff knowledge and familiarity with the resources of other agencies and community groups.

Differentiating services integration from service comprehensiveness. Services integration is intended to provide not only faster, more appropriate services, but also services that would not otherwise be available to certain clients. The referral process educates clients on the options and assistance potentially available. Improved interagency planning and coordination reduces the barriers to obtaining additional services. All this makes the task of differentiating services integration from service comprehensiveness very difficult. Evaluation and program staff need to develop clear expectations on the extent to which the ease of obtaining services and the appropriateness of the service package can be distinguished from the extent to which the program is providing comprehensive services to meet the full range of client needs.

Conclusions

Strong pressure to demonstrate program impacts dictates making evaluation activities a required and intrinsic part of program activities from the start. At the very least, evaluation activities should include performance monitoring. The collection and analysis of data on program progress and process builds the capacity for self-evaluation and contributes to good program management and efforts to obtain support for program continuation—for example, when the funding is serving as "seed" money for a program that is intended, if successful, to continue under local sponsorship. Performance monitoring can be extended to non-experimental evaluation with additional analysis of program records and/or client surveys. These evaluation activities may be conducted either by program staff with research training or by an independent evaluator. In either case, training and technical assistance to support program evaluation efforts will be needed to maintain data quality and assist in appropriate analysis and use of the findings.

There are several strong arguments for evaluation designs that go further in documenting program impact. Only experimental or quasi-experimental designs provide convincing evidence that program funds are well invested, and that the program is making a real difference to the well-being of the population served. These evaluations need to be conducted by experienced researchers and supported by adequate budgets. A good strategy may be implementing small-scale programs to test alternative models of service delivery in settings that will allow a stronger impact evaluation design than is possible in a large scale, national program. Often program evaluation should proceed in stages. The first year of program operations can be devoted to process studies and performance monitoring, the information from which can serve as a basis for more extensive evaluation efforts once operations are running smoothly.

Finally, planning to obtain support for the evaluation at every level—community, program staff, agency leadership and funder—should be extensive. Each of these has a stake in the results. Each should have a voice in planning. And each should perceive clear benefits from the results. Only in this way will the results be acknowledged as valid and actually used for program improvement.

Notes

1. Connell, J.P., Kubisch, A.C., Schorr, L.B., and Weiss, C.H. (1995) *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*. Washington, DC: The Aspen Institute.
2. Kumpfer, K.L, Shur, G.H., Ross, J.H., Bunnell, K.K., Librett, J.J. and Milward, A.R. (1993) *Measurements in Prevention: A Manual on Selecting and Using Instruments to Evaluate Prevention Programs*. Public Health Service, U.S. Department of Health and Human Services, (SMA) 93-2041.
3. For more information on deciding when and how to make decisions on whether and how to conduct a program evaluation, see Schmidt, R.E., J.B. Bell, and JW. Scanlon (1979), "Evaluability Assessment: Making Public Programs Work Better," *Human Services Monograph*

- Series, 14: 4-5.* Washington, DC; and Wholey, Joseph S. (1994), "Assessing the Feasibility and Likely Usefulness of Evaluation." In Joseph S. Wholey, Harry P. Hatry, and Katherine E. Newcomer (eds.), *Handbook of Practical Evaluation, 15-39.* San Francisco: Jossey-Bass.
4. Berk, R.A., and Sherman, L.W. (1988) "Police Responses to Family Violence Incidents: An Analysis of an Experimental Design with Incomplete Randomization." *Journal of the American Statistical Association* 83(401):70-76.
 5. Kalbfleish, J.D., and Prentice, K.L. (1980) *The Statistical Analysis of Failure Time Data.* New York: Wiley.
 6. Rhodes, W.M. (1986) "A Survival Model with Dependent Competing Events and Right-hand Censoring: Probation and Parole as an Illustration." *Journal of Quantitative Criminology* 2(2): 113-138.
 7. Ellickson, P.L., Bell, R.M., and McGuigan, K. (1993) "Preventing Adolescent Drug Use: Long-Term Results of a Junior High School Program." *American Journal of Public Health* 83(6): 856-861.
 8. See Campbell, D.T. and Stanley, J.C. (1963) *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally.
 9. Campbell and Stanley (1963).
 10. Loftin, C., McDowall, D., Wiersma, B., and Cottey, T.J. (1991) "Effects of Restrictive Licensing of Handguns on Homicide and Suicide in the District of Columbia." *New England - journal of Medicine* 325 (December 5): 1615-1620.
 11. Heckman, J.J. (1979) "Sample Selection Bias as a Specification Error." *Econometrica* 47:153-162.
 12. Joreskog, K.G. (1977) "Structural Equation Models in the Social Sciences." In P.R. Krishnaiah (ed.), *Applications of Statistics, 265-287.* Amsterdam: North-Holland; Bryk, A.S. and Raudenbush, S.W. (1992) *Hierarchical Linear Models: Applications and Meta-Analysis Techniques.* Newbury Park, CA: Sage.
 13. Roosa, M.W. and Vaughan, L. (1983) "Teen Mothers Enrolled in an Alternative Parenting Program: A Comparison with Their Peers." *Urban Education* 18: 348-360.
 14. Engle, R-F and Granger, CW.J. (1987) "Cointegration and Error Correction: Representation, Estimation and Testing." *Econometrica* 55: 251-276.
 15. Barker, G. and Fontes, M. (1995) "Review and Analysis of International Experience with Programs Targeted at At-Risk Youth." Paper prepared for the World Bank.
 16. Barker and Fontes (1995).

17. Chaiken, M. (1990) "Evaluation of Girls Clubs of America's Friendly PEERsuasion Program." In R.R. Watson (ed.), *Drug and Alcohol Abuse Prevention*, 265-287. Clifton, NJ: Humana Press.
18. Rossman, S.B and Morley, E. (1994) *The National Evaluation of Cities in Schools*. Report submitted to the Office of Juvenile Justice and Delinquency Prevention. Washington, DC: The Urban Institute.
19. Cohen, M. (1994) "The Monetary Value of Saving a High Risk Youth. " Draft report. Washington, DC: The Urban Institute.
20. Kim, S., Coletti, S.D., Crutchfield, C.C., Williams, C. and Hepler, N. (1995) "Benefit-Cost Analysis of Drug Abuse Prevention Programs: A Macroscopic Approach." *Journal of Drug Education* 25(2): 1 11-127.
21. Burt, M. R. and Resnick, G. (1992) *Youth at Risk: Evaluation Issues*. Washington, DC: The Urban Institute.
22. Barker and Fontes (1995).