There is a great deal of enthusiasm for applying “business principles” and “investment analysis” to decisions about funding early childhood interventions. The “discipline” associated with these hard-nosed business management approaches is perceived to be a useful antidote to the often emotional appeals and political rancor that accompany policy discussions and decisionmaking in this area. Irrespective of one’s view of the relative merits of such methods as cost-benefit analysis for informing policy, cost and outcome methods have emerged as one of the most prevalent tools in the public policy arena (Adler and Posner, 2000). In fact, many states and the federal government have mandated the use of such methods as cost-benefit analysis as part of the policy calculus for various types of policies (Hahn, 2000).

A variety of terms are used, sometimes imprecisely, to refer to the methods in the general class of cost and outcome analyses, including benefit-cost analysis and cost-effectiveness, among others. This chapter will define and illustrate these various concepts and also point out their limitations.¹ We note at the outset, however, that the art and science of quantitative analysis of management problems is far broader than any one of—or even the entire collection of—these notions.²

¹Some useful references for further reading are Gramlich (1981), Keeney and Raiffa (1976), Yates (1996), Mishan (1998), and the June 2000 issue of the Journal of Legal Studies.

²Other tools include the more mathematically advanced methods of operations research, including Monte Carlo simulation, analysis of risk attitudes, Multi-Attribute
GENERAL FRAMEWORK

Over the years, RAND has developed a structured approach for quantitatively analyzing management problems. Called policy analysis or policy scorecard analysis and is specifically intended for issues involving complex systems and competing interest groups (stakeholders) with different and frequently conflicting goals (Quade, 1989). Policy scorecard analysis requires one to take a broad, systems view of a problem. The problem formulation must include a wide enough range of impact measures to reflect the concerns and goals of all the stakeholders and a wide enough range of alternative policies to map the major trade-offs among the impact measures. Policy scorecard analysis has been applied to a variety of issues such as water management (Goeller et al., 1977; Goeller and the Pawn Team, 1985; Walker et al., 1993), air quality (Goeller et al., 1973), transportation (Hillestad et al., 1996; Walker et al., 1999), drug policy (Caulkins et al., 1997; Caulkins et al., 1999), education (Benjamin et al., 1993; Park and Lempert, 1998), and early childhood programs (Karoly et al., 1998).

Policy scorecard analysis provides a framework within which one can employ the cost and outcome methods mentioned above. We will begin by describing policy scorecard analysis and then use the framework to distinguish among the various cost and outcome methods.

The Policy Analysis Scorecard

A central construct in policy scorecard analysis is the scorecard (see Table 2.1). This is simply a table with a column for each policy and a row for each impact measure. Where possible, entries in the table should be cardinal measures of the size of an impact (e.g., policy A
costs $125 million per year). But they may be rankings (policy B is first, followed by A, D, and C in that order) or categories (High versus Low or Good versus Intermediate versus Bad) or even text descriptions (policy A has special feature X).\(^4\) To select the preferred policy, the decisionmaker will compare the columns in the scorecard to determine which one he or she prefers. Typically, no policy will beat all the others on every impact measure, so selecting a policy will involve trading off one impact against another.\(^5\)

At the end of a study, a scorecard is often a good way to summarize the results of an analysis to the sponsor. For this purpose, the analyst must restrict the size of the scorecard, so the scorecard will pre-


\(^5\)The notion of a trade-off can be illustrated as follows. Anyone would agree that it’s better to be rich and healthy than poor and sick! But there may be no way to achieve both objectives simultaneously. One may have to sacrifice some of one to obtain more of the other, for example, by cutting back on work (and hence income) to reduce stress-related disorders.

<table>
<thead>
<tr>
<th>Table 2.1</th>
<th>Illustrative Scorecard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Impacts</strong></td>
<td><strong>Alternative Policies</strong></td>
</tr>
<tr>
<td>Program Descriptors</td>
<td>Parent training</td>
</tr>
<tr>
<td>Cost Elements</td>
<td>Labor paid by agency</td>
</tr>
</tbody>
</table>

\[298x674\]Overview of Cost and Outcome Analysis 7
sent only a handful of alternatives (columns) and impacts (rows) that illustrate the major choices and key trade-offs. At the beginning of a study, constructing a notional scorecard is a useful aid to problem formulation. The major tasks of formulation are specifying the range of alternatives (columns of the scorecard) to be considered, specifying the kinds of impacts (rows of the scorecard) to be estimated, and specifying how those impacts will be measured (entries in the cells of the scorecard). Initial formulation of the problem will typically produce far too many alternatives and impacts to be included in an actual scorecard, and a large part of the analyst's art consists of screening out the less desirable alternatives and the less useful impacts, ending with a scorecard of manageable size that does not mislead the client.

The scorecard is most obviously an appropriate construct for decision problems, such as selecting one program from among several alternatives or designing a program that maximizes the return on investment or that maximizes the effectiveness for a given budget or that minimizes the cost while achieving specified outcomes. Less obviously, the scorecard construct is also appropriate for the task of program evaluation, where at first glance it appears that only one program exists and no alternatives need be considered.

Initial appearances can be deceiving. Most fundamentally, even defining the costs and benefits of a program requires distinguishing what is part of the program from what is not. To say this another way, it requires establishing a baseline, a state of the world without the program that can be compared to the world with the program in place. In clinical trials of a new drug, for example, the baseline is established by a control group of subjects who do not receive the drug. They are compared to subjects similar in all ways except that they are given the drug.

Beyond this, a program is usually evaluated with an eye toward improving it, replicating it in a different setting (e.g., serving a different population), scaling it up, or perhaps canceling it. That is, a program evaluation is generally expected to lead to a decision. A decision to cancel the program will be based on a comparison of the program to the baseline. Improving the program, replicating it, or scaling it up or down will involve comparing the program as currently implemented with one or more variations of the program.
In the remainder of this chapter, we discuss three questions:

- What policies (columns) and impacts (rows) should appear in the scorecard?
- How do we fill out the body of the scorecard?
- Once the scorecard has been constructed and filled out, what methods do we employ to attain our analysis objectives? The methods we will consider are the four listed previously, namely, benefit-cost analysis, cost-savings analysis, cost-effectiveness analysis, and cost analysis alone.

This discussion will proceed linearly, whereas in an actual study the analyst would iterate among these steps. Early in the study an analyst will tentatively select policies to consider but may later discover that information about some policies is simply too difficult to collect, and these policies must be dropped. Or an analyst may discover that none of the policies offer benefits to a particular stakeholder group and try to design a new policy that fills that void. For similar reasons, the analyst may add or delete impacts during the course of the study.

**SELECTING POLICIES AND IMPACTS**

When someone argues that a program or policy is the “best” way (or even a “good” way) to solve problem X (where, for example, X is traffic congestion or air pollution or drug abuse or child neglect), an important reaction should be to ask, “Compared to what?” The columns in the scorecard answer this question. Looking at the scorecard, the analyst and the decisionmaker can compare the policies that exist, but they can only speculate about policies that have been omitted.

A second important question is, “How do you measure the ‘goodness’ of the policy?” Or to say this another way, what are the costs, the products, the side effects, the unintended consequences? The rows in the scorecard answer this question. The analyst and the decisionmaker can consider costs or population served or any other impact only if it is included in the scorecard.

Selecting the rows and columns of the scorecard is thus a key aspect of a study design, with decisions about whether rows and columns
are defined in a more limited fashion or more expansively, along with the specific elements to include in each dimension. We discuss each of these aspects in turn.

**Broad Versus Narrow Formulation of the Problem**

Formulating a policy problem broadly means including a wide range of alternative policies and impacts. A broad formulation has both advantages and disadvantages. Data gathering and analysis for a wide range of policies and impacts will be more costly, time-consuming, and difficult than for a narrow range. If the choice of a preferred policy is to be made by a group, consensus will be harder to achieve when there are many alternatives to choose from and many impacts on which to compare them. On the other hand, a narrow formulation may exclude impacts that measure important costs and benefits and may ignore policies that excel on the excluded impacts. There was a time, for example, when factories were sited, built, and operated without regard for their environmental impacts.

If the objective of the analysis is to improve an existing policy or to replicate a policy in a new environment, it is important that there be adequate variation among the policies in the scorecard. The role of analysis in this context is to do as much policy improvement or policy adaptation on paper (or by computer) as one can, so that the worst features can be weeded out before the policy is actually delivered to real people.

**The Baseline and Alternative Policies**

The illustrative scorecard above includes a column for a policy or program labeled “baseline.” Typically, this policy represents the world without the alternative policies or programs under consideration. For many program evaluations, the baseline is the control group or comparison group. In experimental evaluations, individuals are randomly assigned to the control group (i.e., the group that

---

6 We distinguish here between a wide range versus a large number of policies and impacts. It is possible to inflate the number of policies or impacts by including numerous minor variations of either, but this does not increase the breadth of the formulation.
receives no new program services or faces the status quo) or the treatment group (i.e., the group that receives the program services or faces the policy alternative). When properly implemented, randomized experimental designs are considered the “gold standard” for evaluation research because the control and treatment groups are as similar as possible except for participation in the program. Thus, any differences in the cells of Table 2.1 can be attributed to the impact of the program or policy. Quasiexperimental designs include a comparison group chosen on the basis of matched characteristics but not random assignment.

The column corresponding to the baseline also provides a place to record scenario assumptions, i.e., assumptions about aspects of the future state of the world that may influence the impacts of the other policies. We will have more to say about scenario assumptions later. The overall objective of the analysis is to compare this baseline to columns representing the various alternative programs or policies and assessing which column represents the optimal choice, given the choice mechanism selected. We will return to the discussion of how to choose among alternatives below.

Typically, all policies save the baseline will be constructed by combining policy elements. For example, a policy element may involve delivering a particular service or intervention (e.g., drug counseling or parenting training) to a specified target population (e.g., low-income first-time mothers in a particular neighborhood) by a certain method (e.g., home visits or sessions at a clinic). Then a policy might deliver different services to several different populations (e.g., parenting training to one group, drug counseling to another). It might deliver different services at different venues. Any particular policy will probably have a fairly well-defined service area, which will be the same for all services it delivers and all populations it serves. Different policies can serve different areas, however.

**Considerations in Selecting Impacts**

As seen in Table 2.1, the illustrative scorecard includes rows for program design, as well as those capturing cost elements and outcome measures. Both the cost elements and the outcomes should be broken out by stakeholder and by time. Breaking out costs and outcomes by stakeholder means identifying who pays or benefits. This
is important because the costs and benefits of a program might accrue to different stakeholders, which is likely to enter the decisionmaking process. For example, a policy or program that benefits group A at the expense of group B will often be opposed by the latter, even if total benefits exceed total costs. Breaking out the costs and outcomes by time means specifying when the cost is incurred or the benefit realized. A policy that incurs costs today but yields benefits only years later may not appeal to a term-limited politician, even though the policy might appeal to somebody with a longer view.

Typically, this implies that the scorecard will have a large number of rows. In many problems it is easy to identify half a dozen stakeholders, e.g., the government agency implementing the policy, two or three other agencies, the target population, family members of the targeted population, and other residents. The analyst will define at least one impact for each stakeholder (e.g., cost) and several outcomes for the targeted population. Each impact may occur this year or in any of the next N years. It can add up to dozens or even hundreds of rows.7

Some outcomes may take so long to be realized that they cannot be observed before the decisionmaker must choose a policy. Early childhood interventions are intended, among other things, to reduce the likelihood that the child will drop out of school or use drugs or commit crimes as an adolescent or young adult. A decade or more must pass before we can observe whether these goals have been met. In place of these key but sometimes unobservable impacts, the analyst must substitute short-term outcomes that are reasonable predictors of the more important long-term outcomes. But “reasonable predictors” is a flexible term. It may be that nothing that can be observed within (say) two years has been demonstrated (e.g., by a

7The sheer size of the scorecard should not be a cause for dismay. At initial formulation, the scorecard will include many more impacts (and alternatives) than it will toward the end of the study. A major part of the analyst’s art is devoted to screening out alternatives and impacts that are not informative. Moreover, for presenting final results to the client, the analyst may split the one scorecard into many, each with a different focus. For example, if the focus is on how the state of the world (e.g., the unemployment rate) affects the performance of different programs, the analyst can construct a handful of scenarios (e.g., “pessimistic,” “best guess,” and “optimistic”) and create one scorecard for each. Or if the focus is on performance in the short run versus the long run, the analyst could construct one scorecard with impacts at one year, another with impacts at five years, and so on.
careful clinical trial) to be a “good” (e.g., acceptable by academic standards) predictor of a future outcome. It is better to include a predictor that is deficient by academic standards than to omit the impact from the analysis. As we discuss later, however, the subsequent analysis must take due account of the impact’s uncertainty.

**FILLING THE CELLS IN THE SCORECARD**

To complete the scorecard, the individual cells must be completed. In this section, we offer several guidelines to be followed, as well as methodological issues that arise as part of this process.

**Express Impacts in “Natural” Units**

Entries in the scorecard should be expressed in “natural” units. That is, where possible, they should be cardinal measures of the size of an impact (e.g., policy A costs $125 million per year). However, cardinal measures—those that can be expressed quantitatively in well-defined units—will not always be available. Where necessary, such as for qualitative impacts, entries may be rankings (policy B is first, followed by A, D, and C) or categories (High versus Low or Good versus Intermediate versus Bad) or even descriptions (policy A has feature X). The reason for this advice is that analysis is often criticized for ignoring considerations that cannot be quantified easily (for example, see Sen, 2000). Including difficult-to-quantify impacts (i.e., impacts for which cardinal measures are hard to define) preserves a chance, at least, to include them in the analysis. Even if they can’t be included in the analysis except by artificial and labored means, they can nonetheless figure in the deliberations of the decisionmaker.8

**Record Cost Elements as Resource Quantities, Not Dollars**

In particular, cost elements (one of the categories of impacts shown in the illustrative scorecard) should generally be shown as quantities of resources, such as man-years or gallons of gasoline. They should

---

8Analysis has limitations, after all. The analyst does not replace the decisionmaker. Rather, he or she collects, processes, and displays information in a way that will help the decisionmaker arrive at better decisions.
not be expressed directly as dollars unless the resource inventories behind the dollars are unavailable, even though the analyst intends to price them out later during the analysis phase of the study. There are several reasons for this. First, prices differ from place to place. For example, a program implemented in one city may make use of volunteer labor and donated facilities, while a similar program in another city may need to pay for some or all of these resources. Second, resources may be shared, and a reported dollar cost will be based on accounting assumptions about whose budget is charged for how much of the resource. Those accounting assumptions can differ for a program implemented elsewhere. Third, some resources may be hard to get quickly or even hard to get at all. It might be necessary to find an alternative way to do things in order to implement the program in another location. For example, there may be no emergency room available in a rural setting, while there will be one in a city.

Many Entries May Have to Be Calculated

Entries in the scorecard can come from a variety of sources. The most obvious is direct measurement, either by the analyst or by others (e.g., an experiment or demonstration reported in the literature). Because few policies in the scorecard will have been implemented in their entirety, direct measurements of their impacts will not exist. Data on the impacts of individual policy elements often will exist, however, and just as a policy is built from policy elements, so too can the impacts of a policy be estimated from the impacts of its elements.

A rather simple model will often serve to estimate the resources employed in a program, as a function of its service area, its capacity (i.e., the number of people the program is designed to serve), and its workload (the number it actually serves). Simple geometric arguments can provide estimates of travel distances, which can easily be converted to travel times (at so many miles per hour) and transportation costs. The workload (number of people served) usually translates easily into direct hours of labor (e.g., so many visits per person served times so many minutes per visit plus travel time).9

It is important to add in indirect hours as well. For example, in addition to time spent directly delivering a service, a service provider will also spend time completing paperwork or engaged in other administrative tasks required for direct service deliv-
Likewise, model-based estimates of benefits may be possible when direct observation is not available. In some cases, longer-term impacts may be projected based on short-term outcomes using relationships estimated in other studies or derived from meta-analyses. In the early childhood literature, for instance, estimates of adult lifetime earnings have been projected based on observed final educational attainment or labor market outcomes in early adulthood (see Chapter Four). Ideally, these projections reflect the latest understanding in the literature and will acknowledge the degree of sophistication of the models and their acceptance by other analysts.

Indirect costs and benefits—those tangentially associated with the program or services being evaluated—may also need to be estimated. One example of an indirect cost is an increase in the use of pediatric care by a participant in a program that provides other types of early childhood intervention services. To obtain this from actual measurements, one must measure the use of pediatric care by participants and by a control group, and subtract the second from the first. (Data about the control group will help fill the “baseline” column of the scorecard.) In the absence of actual measurements, one might bound the cost by assuming participants will use pediatric care at whatever rate the American Medical Association recommends.

Because ideal data for each entry in the scorecard are not likely to be available, the analyst must use creativity and informed guesswork to fill it in. Rarely will there be enough data of high enough quality that all entries can be estimated with high confidence. Large blocks of entries may need to be based on educated guesswork if they are not to be left entirely blank. Of course this affects the reliability of the analysis, but in our view, it should not be taken as an excuse to abandon analysis altogether (see the discussion in Quade, 1989).

**Explicitly Address Statistical Uncertainty**

Entries in the scorecard will be uncertain. Some of this uncertainty will be of the familiar statistical variety.\(^{10}\) Survey results will have an

---

\(^{10}\) Another source of potential uncertainty is errors in measurement. Data quality concerns are relevant for both cost and outcome measures, and may be an issue with
error of x percent. Estimates from an equation fitted to data by ordinary least squares will have a standard error. The sizes of these errors should be shown in the scorecard so the analyst and decisionmaker can judge whether two policies differ significantly in a particular impact. When available, these errors can also be used with the aggregation methods discussed below to provide estimates of the uncertainty associated in the cost-benefit, cost-effectiveness, and related analysis. In some cases, only a subjective characterization of uncertainty is available, but even subjective characterizations are generally better than providing only point estimates of quantities that are in fact uncertain.

Moreover, one should distinguish between the statistical significance and practical significance of such a difference. If the standard error of an impact's estimate is low, two policies may have a statistically significant but practically inconsequential difference in that impact. By contrast, if the standard error is large, the difference may be statistically insignificant but practically important. In the latter case, it is not known whether the difference is real, but it is important to find out. One case where the error may be large is when a short-term impact has been used as a predictor of an important long-term outcome.11

The issue of statistical uncertainty means that sample size considerations are important at the design stage of a program evaluation, both for measuring program impacts and for conducting related cost and outcome analyses. Typically, in experimental and quasiexperimental study designs, sample sizes for treatment and comparison/control groups are chosen by balancing cost and other implementation concerns against the statistical power to detect differences between the two groups. If cost and outcome analysis is planned, the decision about sample size will have implications for the ability to draw inferences about program differences in economic terms as well.

information obtained through direct observation, surveys, or administrative sources. Ideally, the most reliable source of data is available for any given scorecard element and any known concerns about data quality are acknowledged by the analyst.

11See discussion in Caulkins et al. (1999) for an illustration of how statistical uncertainty can affect and be handled in cost analysis.
Explicitly Address Scenario Uncertainty

In most cases, factors completely outside the policies of interest will affect the sizes of the impacts. For example, five years ago a family enrolled in an early childhood intervention program could, in principle, remain on public assistance indefinitely. Under current law, the family will be dropped from the rolls after a few years. Depending on the scenario, the eventual benefits of helping a mother, or eventually a child, enter the workforce will be quite different. Or suppose the program provides job training (or refers participants for job training). The effectiveness of this service depends on the local availability of jobs, which in turn depends on the state of the economy. Policymakers should be made aware of assumptions about future developments that may drive the success or failure of the program (Dewar, 1993).

Including a baseline in the scorecard provides a vehicle for including scenario assumptions. Frequently an analyst or decisionmaker will talk about the cost or benefit of a policy or program, with no reference to the baseline at all. This is a convenient shorthand, but it suppresses the fact that the costs and benefits depend on more than the features of a policy or program. They depend as well on the environment in which the policy is implemented and the future environment in which it is operated—for example, the population the program is serving or the other services available in an area. Thus, ideally, the analyst describes the baseline in a rich enough manner that it includes all of the assumptions about the future state of the world that are likely to affect the performance of any of the policies. If the analyst anticipates replicating a policy in another environment, the baseline should also include any factor that may differ between the current and target environments, if that factor influences the performance of any policy.

Account for Time Path of Benefits and Costs by Discounting

A final consideration in filling in the cells of the scorecard involves how to value costs or benefits that accrue in the future. For example, suppose that a home visiting program for 100 children would reduce the expected number of emergency room (ER) visits per child in each of the subsequent three years by one visit. If each ER visit costs an
average of $200, one might think the benefit is best described as the elimination of one visit per year per child: 100 participants x 1 visit per participant x $200 per visit x 3 years = $60,000. But the usual practice is to weight or value outcomes that occur sooner more than outcomes that are delayed. It is obvious why this should be so with money. One would rather have $1,000 today than $1,000 next year, because if a person had $1,000 today he or she could invest it and have more than $1,000 next year. The same logic of “discounting” or applying “time preferences” can be applied to nonmonetary outcomes, and at the same rate (Keeler and Cretin, 1983).

While there is consensus that future outcomes should be discounted, there is no consensus as to what rate should be used, although 4 percent is typical.12 If we apply a 4 percent discount rate to this example, we would calculate the “present value” of reducing ER visits as the amount saved per year, scaled by a discount factor, which is $20,000 + $20,000 x (1/1.04^1) + $20,000 x (1/1.04^2), or $57,700. The term “present value” connotes the idea that given a 4 percent discount rate, one should feel the same about receiving $57,700 today and receiving a savings of $20,000 at the end of each of the next three years. In terms of nonmonetary outcomes, you could discount the 100 ER visits per year for the next three years by the same rate to get a present value of 289 visits. While discounting is a routine method in analysis, to simplify exposition and focus on the more fundamental conceptual issues, it will be suppressed in the remainder of this discussion.

COMPARING POLICIES

Once the analyst has a scorecard with all the cells filled in, it is possible to compare the policies. The purpose of the analysis is likely to be one of the following:

---

12In medicine, 3 percent and 5 percent are recommended (Gold et al., 1996). A variety of RAND analyses in the drug, criminal justice, and children and youth intervention policy areas have used a 4 percent discount rate (e.g., Rydel and Everingham, 1994), while Karoly et al. (1998) explicitly consider a range of discount rates from 0 to 8 percent. Rates between 0 and 10 percent or higher have also been used. The choice of rate may be a function of the time preference of the stakeholder or decisionmaker.
• Select the “best” policy (column) in the scorecard.
• Design a new policy that is “better” than any of the policies in the scorecard.

Since policies have many different impacts, it is highly likely that one will be better than its alternatives on some impacts but worse on others. Comparing policies therefore requires trade-offs to be made among the impacts. Analysts often devise metrics that summarize most or all of the impacts into a single, aggregate score. These metrics define trade-offs among the impacts, because a unit improvement in one impact is worth whatever size reduction in a second impact is necessary to keep the score constant.

Not all methods of selecting a “best” policy use a single aggregate measure of merit. One common method, called a constant-cost analysis, uses one measure of effectiveness and one of cost and deems the policy “best” that maximizes the effectiveness measure while not exceeding a specified cost. If cost is defined from the point of view of the decisionmaker, it is sometimes called a constant-budget analysis. Another method, called a constant-effectiveness analysis, permits the use of several measures of effectiveness and one of cost. The policy is deemed “best” that achieves specified levels of each of the effectiveness measures while minimizing cost. These methods are only useful, however, if they rely on a small number of measures. Thus they require the impacts in the scorecard to be substantially aggregated. We now review some of the alternative ways of creating summary metrics of the costs and benefits of policies.13

**Common Methods for Aggregating Impacts**14

Cost-benefit analysis converts the benefits and costs into common units, most often dollars, and then notes which is greater. Benefits
that cannot be expressed in dollar terms cannot be compared and are excluded from the formal analysis. The purpose of cost-benefit analysis is to help in deciding whether a program is of value to the decisionmaker, or notional decisionmaker, when the analysis is done from the perspective of society at large. The greater the margin by which benefits exceed costs, the better the investment we consider the program to be.15

One distinction among approaches to comparing costs and benefits concerns the stakeholder to whom costs and benefits accrue. Cost savings analysis is a term sometimes used to refer to a cost-benefit analysis done from the perspective of the government generally or a particular government agency. It compares only the costs to government and the savings to government generated from a program. Cost savings analysis is used when asking questions, such as whether the benefits of a program to government pay back the costs taxpayers invested in the program.

The two common ways to compare the benefits and costs are by looking at their ratio or their difference. Dividing the benefits by the costs yields a benefit-cost ratio. Referring to our example of the home visiting program above, suppose the program cost $300 per child, for a total cost of $30,000. Then, the benefit-cost ratio for the program is $57,700/$30,000 or 1.9. Subtracting costs from benefits yields the net value. Because discounting is often involved, this is most often called the net present value, or NPV.16 In our example, the NPV of the parent-training program is $57,700 – $30,000 = $27,700.

When other program alternatives to this treatment program exist, one should generally choose the program with the greatest measure of merit. For example, if three alternative home visiting programs have NPVs of $15,000, $27,700, and $45,000, respectively, and you can only implement one, choose the last. Note, however, that using

---

15It might seem natural to say that if benefits exceed costs, then the program is a good investment. But this ignores the question, “Compared to what?” That is, the question is not whether the investment is “good” in some absolute sense, but whether it is better than the alternatives.

16See Karoly et al. (1998) and Currie (forthcoming) for examples of cost analysis of early childhood programs that use the NPV approach.
the benefit-cost ratio may lead you to choose a different alternative, if the costs of the alternatives are substantially different.

Cost-effectiveness analysis tries to side-step uncertainties about how to value different aspects of programs by looking at the ratio of benefits to costs without reducing them to common units. For example, our hypothetical home visiting program has a cost-effectiveness ratio of 289 ER visits averted / $30,000 in program costs = 9.6 ER visits averted per thousand dollars spent. The ratio of effectiveness to cost is sometimes informally termed the “bang for the buck.” This term comes from cost-effectiveness analysis in the military context, where monetizing outcomes, such as the ability to deliver a given payload of bombs, is similarly difficult. In other contexts, it is common to invert the ratio, calculating the cost per unit of benefit purchased. For instance, health care programs are often evaluated in terms of the cost per quality adjusted life year (QALY) saved (Kamlet, 1992). In those cases, smaller numbers indicate more efficient programs.17 Whether it is more felicitous to think about maximizing what is obtained for a given cost or minimizing the cost necessary to attain a given effect depends on the context. The term cost-effectiveness covers both variants, although calculations of cost per QALY are sometimes called cost-utility analyses.

The cost-effectiveness ratio for a single program is often difficult to interpret. Most people do not have an intuitive sense of whether averting 9.6 ER visits per thousand dollars is a lot or a little. But if one calculates the cost-effectiveness ratio for each available intervention, the one with the highest ratio is the preferred place to invest the next dollar. (If the ratios are computed in terms of cost per unit benefit, not benefit per unit cost, then the intervention with the smallest ratio would be preferred.) For example, if alternatives to the home visiting program had cost-effectiveness ratios varying between two and seven ER visits per thousand dollars spent, then the home visiting program would, all other things being equal, seem to be a more appealing place to invest the next thousand dollars.

One can also compare programs in terms of the lengths of time they must remain in operation to recoup the initial investment, some-

---

17 See Greenwood et al. (1998) and Caulkins et al. (1999) for examples of cost-effectiveness analysis for early childhood programs.
times called the payback period. Typically, for a given treatment population in the early stages of a program, only costs are generated. Once the program services end for that population, the cumulative costs do not change. During the period of program implementation, benefits may begin to accrue and they can continue to grow after the program services end. For example, Karoly et al. (1998) found that the Elmira home visiting program paid back its costs of delivering services to the treatment group after about two years, while the Perry Preschool Program took nearly two decades to recoup its costs for the cohort it served.

As discussed above, programs often produce multiple benefits. For example, a substance abuse treatment program might not only reduce cocaine use, but it might also avert a given number of serious crimes and the years of prison time associated with those crimes. Cost-effectiveness ratios per se are limited to a single outcome and so have a hard time fully reflecting such a range of benefits. But, sometimes the candidate interventions produce the various benefits in almost fixed proportions. In that case, focusing on one benefit is not problematic because whichever program generates the most “bang for the buck” with respect to that benefit does so with respect to the other benefits as well. But that is by no means always the case. For example, drug prevention programs reduce the number of cocaine users by a greater proportion than they reduce the quantity consumed; for drug treatment programs, the opposite is true.

When outcomes are produced in different proportions, one may calculate a cost-effectiveness ratio for each important outcome. This is sometimes called cost-consequences analysis. For some purposes listing explicitly the set of outcomes produced per thousand or per million dollars invested is useful. For others, decisionmakers may prefer a single, bottom-line summary. Cost-benefit analysis provides that bottom-line summary by reducing all outcomes to a common currency.

---

18 For example, in Greenwood et al. (1998) incarceration policies tended to produce reductions in different types of crime in constant proportions, so the analysis could usefully focus on one aggregate measure (serious crimes) without worrying about the fact that some types of serious crime (murder) are in some sense more “costly” per offense than are other serious crimes (e.g., robbery).

19 For other such examples, see Caulkins (2000).
Notice that when some of the benefits are avoided costs, as in the example of reduced crime and use of the criminal justice system, ambiguity can arise with respect to the computation of the benefit-cost ratio. The NPV is the same whether the savings in prison costs are counted as a benefit or a cost offset. But the benefit-cost ratio changes. If one counts the taxpayer savings from reduced prison time as a benefit, the benefit-cost ratio will include the prison cost savings in the numerator. If one views it as a cost offset, it is possible that the net cost to the taxpayers of funding the treatment program is zero or even negative (depending on the size of the offset). Thus, it is possible for the benefit-cost ratio to become negative or to be undefined (e.g., when net costs are zero).20

That one can compute different benefit-cost ratios depending on whether some outcomes are viewed as benefits or cost offsets leads some observers to recommend focusing on the NPV, not the benefit-cost ratio. However, the NPV may depend on the scale of the project. A mediocre program implemented throughout a large state such as California may have a larger NPV than an outstanding program implemented in a small state. In these contexts, it is thus useful to discuss the NPV per unit of activity, such as the NPV per child or family in a program.

Because monetized physical outcomes are not the same as “real” money, one can make an argument for putting all outcomes that literally involve dollars in the denominator and segregating the “dollar equivalent valuations” in the numerator of the benefit-cost ratio. This approach is sometimes labeled cost-offset analysis. When the alternative algorithms suggest different results, the differences should be highlighted and explained.

Finally, cost analysis alone, with no accounting for program benefits, can also be useful to decisionmakers for a variety of purposes—for example, discovering which factors need to be considered in replicating a program elsewhere. Compared with a cost-benefit analysis or the related methods that also require measurement and analysis

---

20 Why would one ever consider avoided costs to be negative costs rather than positive benefits? Because an impact can be negative for policy A and positive for policy B. Whether it is categorized as a cost or a benefit, it will be negative for one policy and positive for the other.
Assessing Costs and Benefits of Early Childhood Interventions

of program benefits, this approach requires the fewest resources to implement, albeit with a corresponding reduction in what is learned about the program’s impacts. It is most valuable when it identifies who bears which portion of the costs, not just the total cost.

Aggregating Impacts Has Disadvantages

Cost-benefit analysis and the allied methods described above collapse the impacts to a single measure of merit, but policymakers answer to the concerns of particular constituencies—perhaps voters, heads of their agencies, clients of their agencies, and others. Also, for most people, decisions are guided by equity and justice as well as efficiency considerations. In short, distributional issues matter.\textsuperscript{21}

If we all agreed how the costs and benefits ought to be distributed among stakeholders, these issues could be incorporated into cost-benefit analyses. One could call improving the lot of criminals a cost rather than a benefit and assign some dollar-equivalent penalty to it. One could decide that from society’s perspective, increasing the income of poor people is worth twice as much per dollar as increasing the income of people in the middle class. One could count as an objective not just improving the average lot of people in different neighborhoods, but also reducing the inequity between them (see, for example, Keeney and Raiffa, 1976). But people differ on these matters, so they place different relative values on various outcomes. As a result, different people will rank policies in different orders, and no single measure of merit will satisfy everybody.\textsuperscript{22}

For example, it might be less costly to implement a publicly funded daycare program in a middle-class neighborhood than it is in a poor

\textsuperscript{21}See Posner (2000), Frank (2000), and Richardson (2000), for discussions of distributional issues for cost and outcome analyses.

\textsuperscript{22}Of course, if the policy choice were up to a single decisionmaker, he or she would use a measure of merit that reflected his or her views, and a suitably tailored cost-benefit analysis would suffice. Policy choices in the real world are often the product of commitments by a range of individuals and institutions. A famous theorem by Kenneth Arrow (1951) demonstrates, roughly speaking, that there is no analytically defensible way to combine the different preference schemes of multiple individuals to obtain a group preference. Thus, different equally justifiable methods of combining individual preferences can lead to different group preferences. Coming to a consensus, therefore, has to be essentially a political process rather than an analytic one.
neighborhood, perhaps because it is easier to find buildings that meet asbestos standards or because fewer of the children have special needs. Furthermore, the impact on tax revenues may be more favorable if the middle-class parents who are freed to work would earn more and be taxed at higher marginal rates than the parents in poor neighborhoods would be. Nevertheless, few would openly sanction targeting such government subsidies at privileged rather than at at-risk families.

A rather infamous example pertains to the value of stolen property. One school (Cook, 1983, and Harwood, Fountain, and Livermore, 1998, are examples) argues that when goods are stolen but not damaged no net loss to society occurs. Society has just as much wealth after the burglary as it did before. The wealth has simply been transferred from one individual to another. Both are members of the society, so there is no net loss. Others (e.g., Trumbell, 1990, and Cohen, 2000) exclude the private gains of criminals, and so view the theft as a loss. Likewise, Cohen would not count the suffering of people incarcerated in a cost-benefit analysis because they are criminals, while others would (Greenberg, 1990).

Even when people agree about the objectives, they may disagree about their priorities. In the case of drug treatment, one person may believe the social costs per gram of cocaine consumed, per serious crime, and per year of incarceration are $100, $10,000, and $25,000, respectively. Another might view drug use per se as less of a problem but believe that crime and incarceration carry hidden costs not reflected in budget-based estimates (e.g., fear of crime spurring middle class flight to the suburbs or the disenfranchisement of minority males by disproportionate rates of incarceration). Inasmuch as estimates of social costs reflect value statements, there is ample room for reasonable people to disagree about the relative costs of various outcomes and, hence, the relative desirability of various interventions.

A related problem stems from differences in opinion about the likelihood of different outcomes. Policy analyses of long-range social investments are fraught with uncertainties, many of which cannot be definitively characterized with objective, historical data. That is not a problem when there is a single decisionmaker. The methods allow and indeed even invite the inclusion of judgment in the form of
“subjective probability assessments.” But when many decisionmakers each have their own personal judgments about not only the likelihood of different outcomes but also the appropriate structuring of the problem, it is much harder for any single report or analysis to guide them collectively.

The result is that benefit-cost studies are sometimes performed from the perspective of a mythical “social planner,” but they are read and judged by individuals with different agendas and different worldviews. A hypothetical early childhood intervention that is cost-justified by its effect on participants’ crime rates a decade or more later when they are adults might not receive the support it “deserves” if the crime declines will bring rewards to the next generation of police commanders, rather than the current generation of social service agency heads, some of whom may not even think of crime prevention as the natural frame for evaluating the programs they sponsor.

Given these concerns, it is important to keep cost-benefit analysis, cost-savings analysis, and other forms of cost and outcome analysis in their place. They can provide valuable input to choosing among different programs, demonstrating a program’s worth, improving programs, and replicating them. But they have their limitations. In any decision, some considerations can be resolved only through a decisionmaker’s values and subjective judgment or through political interaction among stakeholders (Frank, 2000, Posner, 2000, and Richardson, 2000).