7

Effectiveness

Objectives

1. Define measures of effectiveness.
2. Identify valid measures of effectiveness for cost-effectiveness (CE) analysis.
3. Assess the main identification strategies in relation to CE analysis.
4. Describe multiattribute utility functions to enumerate effectiveness.

Economic evaluation must consider both the costs and results of interventions. By comparing both costs and results among alternatives, one can choose the alternative that provides the best results for any given cost outlay or that minimizes the cost for any given result. In previous chapters, the assessment of costs and their measurement were presented. In this chapter, we discuss the effectiveness of educational interventions. In the subsequent chapter, we combine this information with costs in order to evaluate the overall CE of interventions.

Often, when we propose CE analysis of an educational intervention, the immediate question is this: How will you measure effectiveness? The simple answer is "in the same way that any impact evaluator in the social sciences would measure effectiveness." That is, we would use the same measure as chosen by the impact evaluator and apply the same method to identify effects as the impact evaluator would.

In this sense, this chapter will appear to be a case of déjà vu for evaluators, for the heart of the evaluation exercise is often precisely that of ascertaining the effects of interventions on particular criteria. For example, evaluators often face situations in which they are asked to ascertain the impact of alternative curricula on reading scores or the effects of an in-service teacher training program on teacher performance. In this respect, the evaluation of outcomes is a familiar endeavor, and it is not the purpose of this book to provide an exhaustive description of evaluation designs. For this, the reader is advised to consult one of the excellent manuals on evaluation, research design, and econometric identification strategies that already exist (e.g., Angrist & Pischke, 2009; Murnane & Willett, 2010; Newcomer, Hatry, & Wholey, 2015; Rossi, Lipsey, & Freeman, 2004).

Here, our purpose is more specific: It is to consider effectiveness measures that can be used for CE analysis. Economic evaluation puts a heightened emphasize on getting the "right" measure of effectiveness. Fundamentally, any measure of effectiveness should fully reflect the objectives of the intervention so that a valid comparison can be made between the intervention and the counterfactual. Of course, this is also expected for measures that are applied in impact evaluations. For CE, however, effectiveness must be represented by a single number so that it can be expressed as the denominator in a ratio. This is a significant constraint and may shape the outcome measures selected; so we specify some features of effectiveness measures that make them preferred for CE analysis. Next, we review the main ways to identify effects: experimental, quasi-experimental, and correlational. This review is intended to help the analyst ascertain how suitable each identification strategy is for CE analysis. Finally, we describe multiattribute utility functions, a general method by which educational outcomes can be represented in a single number that reflects the objectives of the decisionmaker. Although utility functions are rarely applied in education research, it is important to develop the analysis to allow for policy preferences to be modelled (Chandra, Jena, & Skinner, 2011).

7.1. SPECIFYING EFFECTIVENESS

7.1.1. Examples of Effectiveness Measures

In principle, any measure of effectiveness can be used for CE analysis. Indeed, one advantage of CE analysis is that it is broadly applicable across many areas of education research.

Examples of interventions and their respective effectiveness measures for selected CE analyses are given in Table 7.1. These are U.S. examples; CE studies in developing countries often focus on student achievement or years of schooling as their effectiveness measure (Dhaliwal, Duflo, Glennister, & Tulloch, 2012; McEwan, 2012). As shown in Table 7.1, the range of outcomes available for CE analysis is wide.

As well as the various measures listed in Table 7.1, effectiveness might be counted as the number of reported disciplinary problems, the number of graduates or trainees placed in jobs, or the number of students who complete college. Most CE studies use measures of academic achievement to indicate effectiveness. But studies often use different scales. For ease of comparability, Harris (2009) recommended the use of Cohen's effect size when evaluating interventions using CE analysis, although these effect sizes must be measured in exactly the same way to be comparable across interventions. As discussed next, test scores might be considered the most analytically tractable measures of educational effectiveness, and they are at least a general measure across all students. However, this by no means implies that CE analysis is restricted to such measures. In fact, some tests may

Table 7.1 Examples of Effectiveness Measures for Cost-Effectiveness Analysis

Study	Intervention/Policy	Effectiveness Measure
Levin, Glass, and Meister (1987)	Peer tutoring	Mathematics and reading achievement
Hartman and Fay (1996)	Referral services	Receipt of special education services
Wang et al. (2008)	After-school program (third grade)	Obesity (percentage of body fat)
Yeh (2010)	Teacher board certification	Student achievement
Borman and Hewes (2002)	Success for All	Reading and mathematics scores
Hollands et al. (2014)	Job Corps	High school dropout rate
Bowden and Belfield (2015)	Talent Search program	College access
Hollands et al. (2016)	Wilson Reading System (third grade)	Alphabetics literacy domain

have poor construct validity (e.g., when students score highly on a test but cannot perform the respective competencies; McEwan, 2015). Nevertheless, in principle, any effectiveness measure might be used for CE analysis.

7.1.2. Linking Objectives and Effectiveness

CE analysis is comparative: It involves evaluating one intervention against another or one intervention against the status quo. It is therefore essential that we are comparing apples to apples such that the two interventions are genuine alternatives and can legitimately be ranked or compared based on the selected measure of effectiveness.

For comparability, the measure of effectiveness chosen should reflect as closely as possible the main objective of the alternatives. For example, programs designed to increase reading achievement should select an appropriate reading test as a measure of effectiveness (see Hollands et al., 2016). Dropout prevention programs should be evaluated according to the numbers of potential dropouts that are averted or students who complete each grade. The effectiveness of various physical education programs could be evaluated in terms of the measured improvements that they bring about in the specific physical skills of participants. The measure should therefore be sufficiently comprehensive as to cover all relevant dimensions (e.g., speed and dexterity if both are impacted). This is challenging because many educational interventions have diverse outcomes and do not meet all goals in the same way. For early literacy interventions, for example, effectiveness should capture all facets of literacy, including comprehension, alphabetics and fluency (National Institute of Child Health and Human Development [NICHD], 2000). For socioemotional learning interventions, effectiveness on any specific dimension might need to incorporate all significant changes in behavior, attitudes, or conduct (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011).

Programs with different objectives will have entirely different indicators of effectiveness. So they cannot be readily compared within the CE framework (or even within a relative effectiveness framework). We cannot, for example, use CE analysis to compare the CE of a dropout prevention program and a physical education program. Likewise, we cannot compare a literacy program that focuses on alphabetics with one that focuses on comprehension. (An alternative might be to convert the effects into pecuniary terms and apply benefit-cost [BC] analysis as described in Chapters 9 and 10.) CE analysis emphasizes comparability across interventions. As such, it is essential that outcomes from separate evaluations are equivalent. This equivalence is often very hard to obtain: Studies vary in the constructs used in measurement, and even if the construct is similar, the measurement scale may not be. In their review of What Works Clearinghouse–approved studies of literacy programs, Hollands et al. (2016) identified 32 with positive effects. However, early literacy outcomes were grouped under four different domains (alphabetics, fluency, comprehension, and general reading achievement), each of which included multiple subcategories; few studies used the same scales (e.g., Dynamic Indicators of Basic Early Literacy Skills, or DIBELS).

Effectiveness measures can be intermediate proxies for final outcomes. But these too should be directly related to some program objective (Weiss, Bloom, & Brock, 2014). For many reasons, often as simple as lack of data, evaluators can obtain measures of intermediate outcomes only. For example, Tatto, Nielsen, and Cummings (1991) compared the CE of three Sri Lankan teacher-training programs in raising an intermediate outcome—teacher mastery of subject matter and pedagogy—even as the ultimate goal was improvement in classroom teaching and student learning. Similarly, Hartman and Fay (1996) compared two different methods of referring children to special education in Pennsylvania. The effect was the number of children who received certain kinds of intervention services; the presumption was that these services would ameliorate learning difficulties. Indeed, even many "final" outcomes in education may simply be intermediate ones. For example, academic achievement is often not valued as an end in itself but is valued for its supposed influence either on wages or an increased capacity to participate in a democratic society.

Effectiveness measures should take account of when the effects occur (Harris, 2009). An intervention that is intended to rapidly increase test scores, for example, is preferable to an intervention that does so to the same magnitude but more slowly. Therefore, as with costs, the effects should be discounted back to a present value.

This issue is best illustrated with an example. Imagine that we are conducting CE analysis of three approaches to dropout prevention in high schools that are implemented over a period of 5 years. The measure of effectiveness is the number of dropouts averted in a given year by each program. The programs yield the same number of undiscounted dropouts but at different times: Alternative A yields 100 fewer dropouts in the first year but none in Years 2 through 5. Alternative B yields 20 fewer dropouts each year. Alternative C yields all its 100 fewer dropouts at the end of the fifth year. If effects are not discounted, then the alternatives are judged to be equally effective because they each reduce dropouts by 100. Clearly, in terms of effectiveness Alternative A is the most attractive, and Alternative C is the least attractive. The valid measure of effectiveness is the discounted dropout rate. This is calculated using the present value formula described in earlier chapters. For this example, if the discount rate is 5%:

$$PV_A = \sum_{t\text{-}1}^{n} \frac{\mathrm{E}_t}{(1+r)^{t-1}} = \frac{100}{(1+0.05)^0} = 100$$

$$PV_B = \sum_{t-1}^{n} \frac{\mathrm{E}_t}{(1+r)^{t-1}} = \frac{20}{(1+0.05)^0} + \frac{20}{(1+0.05)^1} + \frac{20}{(1+0.05)^2} + \frac{20}{(1+0.05)^3} + \frac{20}{(1+0.05)^4} = 91$$

$$PV_C = \sum_{t\text{-}1}^{n} \frac{\mathrm{E}_t}{(1+r)^{t-1}} = \frac{100}{(1+0.05)^4} = 82$$

With discounting, Alternative A is 10% more effective relative to Program B, and it is 22% more effective than Program C.

Discount factors for interventions delivered in specific grades and of given durations are reported in Harris (2009, Tables 1 and 2). However, educational CE analyses that discount their effects are rare (for an exception, see Caulkins, Rydell, Everingham, Chisea, & Bushways, 1999; for health literature, where discounting effect is more common, see Weinstein, Torrance, & McGuire, 2009). In general, this is because effects are often measured only at a single point in time (e.g., percentage completing high school) rather than longitudinally or cumulatively. So, evaluators know the percentage of students who did not complete high school when they are aged 18 but not when those students dropped out. Moreover, it is not clear what discount rate to apply. The discount rate for effects need not be the same as for costs. Some health researchers have argued that outcomes should be discounted at a low rate so that policy decisions are not skewed away from preventive interventions (Brouwer, Niessen, Postma, & Rutten, 2005).

Overall, the primary requirement is that measured effectiveness should accurately reflect the objectives of the intervention. This is salient for any impact evaluation. Measures should be valid—that is, bearing a close correspondence to the underlying concept that they are intended to reflect, and they should be reliable, yielding the same results when applied on repeated occasions to the same groups. In principle, any idiosyncratic measure that fully reflects the objectives of the program and offers sufficient comparisons for decisionmaking purposes can be justified for CE analysis.

7.1.3. Single Measures of Effectiveness

As well as matching to objectives, a particular requirement of CE analysis is that the effects of an educational intervention are represented by a single measure. The measure needs to be singular so that it can be readily expressed as the denominator in the CE ratio.

This requirement is reasonable if the alternatives have a single objective. Furthermore, there should be no compelling reason to believe that secondary effects will be produced in other areas—either intentionally or unintentionally. Of course, these assumptions are often unrealistic. Most educational alternatives jointly produce a wide range of outcomes that require numerous measures of effectiveness. For example, we may wish to compare these school investments: lengthening the day in elementary schools and lowering the class size. Lengthening the school day might improve test scores and increase physical activity by students; lowering class size might improve test scores and improve teacher satisfaction. Expressing these effects in a single metric is very challenging. More emphatically, socioemotional learning interventions have been found to influence social skills, attitudes, positive social behavior, and student conduct as well as academic achievement (Durlak et al., 2011; Sklad, Diekstra, De Ritter, Ben, & Gravestein, 2012). To perform CE analysis, it must be valid to represent outcomes for these interventions in a single variable.

The issue of multiple outcomes is important even when the stated objectives of programs have a limited scope. One could imagine three separate programs, each focused on raising the English competencies of recent immigrants. To varying degrees, each program removes children from their standard classroom environments for part of the school day. The fact that children are deprived of some classroom instruction may yield effects (perhaps negative) in other areas, even if the programs succeed in improving English skills. In each of these cases, it behooves the evaluator to measure the important intended and unintended outcomes of each alternative.

Faced with multiple outcomes, the effectiveness measure may be a weighted combination of expected probabilities. For example, an intervention may be delivered to students who want to complete a BA and are deciding on whether to start at community college (see Agan, 2014; Bailey, Jaggars, & Jenkins, 2015). Students can start directly at a four-year college, and their BA degree completion rate can be estimated. Alternatively, students can start at community college; only a proportion of these students will transfer, and a subset of these will ultimately graduate. Their BA completion rate is therefore a product of the transfer and completion probabilities. Hence, we can calculate the expected outcomes of the college-choice intervention as a set of weighted probabilities given the pathways students take through the postsecondary system (Shapiro et al., 2015).

If there is no straightforward way to combine multiple outcomes into a single measure, there are several ways of proceeding. First, the evaluator could conduct a separate CE analysis for each measure of effectiveness. Such an analysis may reveal that a given alternative is to be preferred unambiguously by virtue of its consistently superior CE across many measures. That is, it may yield a given amount of mathematics achievement at a lower cost than other alternatives, and it may also yield a given amount of reading achievement at the lowest cost. In these cases, the evidence clearly supports the use of a particular alternative.

However, it is possible that one alternative is the most cost-effective means of raising mathematics achievement, whereas another is more cost-effective at improving reading. In such an instance, the evaluator could simply present the results of each CE analysis and clearly describe the relevant trade-offs. This approach can also be used if the intervention is differentially effective (e.g., CE results can be presented by sex, race, or socioeconomic status [SES]). In one CE analysis, Levin and colleagues (1987) compared the costs and effects of four interventions: (1) peer tutoring, (2) computer-assisted instruction, (3) class size reduction, and (4) increase in instructional time. The analysis revealed that peer tutoring was the least costly method of obtaining gains in mathematics achievement. While peer tutoring was also the most cost-effective means of raising reading achievement, the analysis showed that computer-assisted instruction assumed the second place. Individual decisionmakers might use these data to make different investment decisions, depending on their priorities.

Alternatively, the evaluator may wish to conduct cost-utility (CU) analysis. In the CU framework, multiple measures of effectiveness—weighted by their importance to parents, administrators, or another audience—are combined into a single summary measure of utility, a decisionmaker’s subjective assessment of value. The weights can be estimated subjectively; if so, the evaluator should consult key stakeholders and carefully consider the primary audience for the analysis. Using more rigorous methods, they could also be elicited from key stakeholders using a formal, structured questionnaire. Such an analysis might reveal, for example, that parents in a particular school district place somewhat higher weight on mathematics achievement than other outcomes. An explication of utility functions and the creation of single effectiveness measures (amounts of utility) is given next.

Clearly, the requirement that only a single measure of effectiveness be applied is a significant restriction on the application of CE analysis. (Again, we note here that this requirement does not apply to BC analysis as it uses dollars as its unit of account. As dollars are fungible and additive, any impacts can be included as long as they can be expressed in dollars. See Chapters 9 and 10.) An intervention that has multiple effects cannot always be reduced to a single number. If so, CE analysis is not appropriate.

Yet, there is a danger in evaluating educational interventions using too many outcomes. It is not untypical for evaluations to claim that their intervention will promote a vector of outcomes, to include child health, socioemotional development, approaches to learning, language development, cognition, and general knowledge (e.g., the Parenting for Life Early Childhood Intervention). An evaluation of this comprehensive nature may yield highly valuable information about specific moderators and mechanisms for enhancing child development. However, it may be difficult for a decisionmaker, when presented with many statistical significance tests across the outcomes, to adjudicate between this program and an alternative early childhood intervention. The advantage of Occam's razor is that a simple claim—parenting programs are more cost-effective than professional development for teachers at improving child health—is easier to test, therefore easier to refute or accept, and hence easier to convey to a decisionmaker.

7.1.4. Appraising Effectiveness Measures

Ultimately, we cannot prescribe the best measure of effectiveness (or as to how many domains that measure is derived from). The appropriate choice will depend on the particular intervention being evaluated and its specific goals. Nevertheless, there are some attributes that make some effectiveness measures preferred over others.

Ideally, an effectiveness measure should be in a ratio scale and continuous. An example here would be a math test where students can score from 0 to 100 and a score of 80 is considered twice as effective as a score of 40 (and 4 times as effective as a score of 20). In this case, it would not matter whether the intervention increases test scores from 20 to 30 or from 80 to 90; both represent a 10-point gain. Similarly, for community college students, most of whom never complete a degree or certificate, the effectiveness measure might be number of credits accumulated (Bailey et al., 2015). For these students, 24 credits are twice as good as 12 credits and 1 better than 23 credits. In developing countries, years of attainment might be valid: Attending school for 10 years may be regarded as twice as effective as 5 years. Nevertheless, the evaluator might propose that getting each student to complete primary school (6 years of attainment per child) is more important than raising the total stock of attainment (e.g., with half the children with 4 years of attainment and the other half having 8 years).

Although test scores may be useful, they need careful interpretation for use with CE analysis. A thorough discussion of construct validity has been provided by Bloom, Hill, Black, and Lipsey (2008) and Lipsey and colleagues (2012). These authors made several points. First, it is necessary to report not only the posttest gain in achievement from an intervention but also the gain relative to the baseline or counterfactual. An intervention may generate a posttest gain even if scores are below baseline; an alternative intervention may generate a smaller posttest gain but all posttest scores are above baseline. In the former case, all scores are going down; in the latter case, all scores are going up. Second, measured gains in achievement may attenuate dramatically over school grades: from first to second grade, effect size gains in math are typically 1; from 11th to 12th grade, effect size gains in math are almost zero (Lipsey et al., 2012, Table 5). An effect size gain of 0.25 in math is modest for a first grader but enormous for an eleventh grader. Finally, there are many different achievement measures and the choice of measure will depend on context.

For CE studies, the analyst might prefer to work with grade equivalent (GE) scores (Lipsey et al., 2012, pp. 23–24). These scores indicate the level of a given student's achievement: 5.3 means the student's achievement is equivalent to that of a student who has completed 3 months of fifth grade. For example, a class-size reduction policy that moves each student from an expected 5.3 GE score to a 5.6 GE score has therefore generated 3 months of achievement per student. A decisionmaker might be able to equate—albeit approximately—this metric to a resource measure for a year of schooling. So, if the school spends $12,000 on academic instruction to move each fifth grader 9 months ahead across all subjects, the decisionmaker might only be interested in class-size reduction if it requires less than $4,000.

By contrast with test scores or other continuous scales, binary or discrete indicators of effectiveness may be less clear for policy purposes. For example, with a mentoring intervention to increase the high school graduation rate, many students will receive resources even though they would have graduated anyway; also, many students will receive resources but still never graduate. Both these groups of students will be considered as zero effect if the effectiveness measure is the number of new graduates. Only the new marginal graduates will be counted in the effectiveness measure. Nevertheless, it is likely that the anyway-graduates and never-graduates received some benefits from the program, even as these benefits are not counted. Expressed as odds ratios, binary indicators can be especially hard to interpret. For example, an odds ratio of 1.22—that is, an increase in the odds of high school completion of 22%—will involve a substantial number of new graduates if the baseline completion rate is 70%. But there will be only a few new graduates if the baseline completion rate is 20%.

These issues gain salience because all resources for all students are counted in CE analysis (we discuss an example in Chapter 9). Imagine a high school mentoring program delivered to 100 students with an impact evaluation that shows the high school completion rate increases from 50% to 60%. This yields 10 new graduates. But the intervention has allocated resources to 100 students of which 90 are counted as a zero effect (the 50 students who would have graduated anyway and the 40 who never graduated). A similar logic applies to the second example where effectiveness is measured using the change in the odds. With CE analysis applied to binary outcomes, it matters a lot how well targeted the intervention is or the prevalence of dropping out. If the mentoring program had been delivered only to the 50 students who were expected to drop out, then the cost is approximately half as large (depending on economies of scale). Therefore, the CE results will be significantly different for programs with different baseline prevalences.

In conclusion, if the program objectives warrant it, the evaluator could choose from a wide range of effectiveness measures. Fundamentally, the CE analysis should use as its effectiveness measure the construct that best captures the goals of the intervention.

7.2. METHODS FOR IDENTIFYING EFFECTIVENESS ●

7.2.1. Experiments, Quasi-Experiments, and Correlational Evaluations

Once measures of effectiveness are established, the next task is to determine whether a particular intervention is successful in altering success on these measures. In particular, we need to ascertain whether there is a cause-and-effect relationship between each alternative and the measure of effectiveness. Does reducing class size lead to increased mathematics and reading achievement? Does an after-school program reduce the likelihood of aggressive behavior?

Typically, this involves comparing the measure of effectiveness for a group of individuals who have been "treated" by the alternative with that of a control or comparison group. There is a vast array of evaluation designs or identification strategies for carrying out these comparisons. For our purposes, we distinguish three categories of evaluation designs: (1) experimental, (2) quasi-experimental, and (3) correlational.

The experimental method directly assigns subjects to a control group and one or more treatment groups. Members of the control group do not participate in the educational alternative that is being evaluated; instead, they provide a baseline estimate of what the treatment group *would have* attained in the absence of the treatment. Ultimately, the estimates of effectiveness are based on the difference between the measured outcomes of the treatment and control groups subsequent to the application of the educational program or policy to the treatment group. Subjects are randomly assigned to control and treatment status. Hence, the groups were equivalent at some initial point and any subsequent difference in outcomes can be causally attributed to the treatment.

Random assignment is a prerequisite of experimental research and is the best way of ensuring equivalence between control and treatment groups. (For a full discussion, see Cook et al. [2002], McEwan [2015], and Smith and Glass [1987]; for the application of experimental methods in health sciences CE analysis, see Greenberg, Rosen, Wacht, Palmer, and Neumann [2010] and Neumann, Greenberg, Ochanski, Stone, and Rosen [2015].) Because randomization provides assurances that the two groups are equivalent on average, prior to the application of the treatment, we can rule out the important threat to validity of group nonequivalence or selection bias. Experiments in education research are also growing in prevalence across a range of educational interventions. A very prominent experiment was on class size (Mosteller, 1995); more recent experiments have tested the effectiveness of guidance programs for college students (Butcher & Visher, 2013), of coaching (Bettinger & Baker, 2014), and of incentive payments in college (Barrow, Schanzenbach, & Claessens, 2015).

The quasi-experimental method relies on plausibly random differences in exposure to an intervention to identify the impact of that intervention. These differences in exposure may be identified using local randomization regression-discontinuity designs, which rely on local randomization of a continuous assignment variable, or instrumental variables designs, which rely on exogenous differences in access to the intervention (see Schlotter, Schwerdt, & Woessman, 2011). Where quasi-experimental methods are not based on random assignment to an intervention, they may still be subject to some selectivity or endogeneity bias (see Heckman & Urzua, 2010; Imbens, 2010). However, evidence from quasi-experimental methods has grown rapidly over recent years (Angrist & Pischke, 2009). Prominent examples include identification of the effects of class size, using exogenous variation of school rules or cohort populations (Angrist & Krueger, 1999; Hoxby, 2000); extra schooling, using exogenous variation in compulsory schooling laws (Oreopoulos, 2006); returns to college based on distance from the local institution (Kane & Rouse, 1995); and the capitalized value of high-quality schooling from additional information (Figlio & Lucas, 2004).

Finally, the correlational method is based on regression analysis controlling for covariates or less restrictive matching estimators (see Imbens & Wooldridge, 2009). When comparing students who receive and do not receive a given "treatment" (e.g., more textbooks or a new curriculum), we make statistical controls for measured characteristics of students (such as a pretest score or SES). If the controls are complete and accurate, then the threat of group nonequivalency is adequately ruled out, but there are often many potential nonobservable influences that make this assumption shaky.

Correlational methods are relatively simple to apply across many educational research topics. Unfortunately, even with matching estimators, it is a fairly tall order to make complete and accurate controls: There are countless unobserved student characteristics that also affect outcomes, such as motivation, family wealth, or ability. If students who received a given treatment tend to possess more or less of these characteristics, then it is quite difficult to separate treatment effects from the preexisting student differences. Statistical procedures can address the threat of group nonequivalence (see Greene, 1997; Wooldridge, 2000), including procedures that specifically model the process of selection into the program or bounding the bias from unobservable characteristics (e.g., Altonji, Elder, & Taber, 2005). But these procedures may not always or completely mitigate bias. Nevertheless, at least since the Coleman Report of 1966 (Coleman et al., 1966), thousands of studies have used nonexperimental data and multiple regression analysis to evaluate educational research interventions.[1] Of particular interest for CE analysis are "production function" studies (Choi, Moon, & Ridder, 2014). These studies attempt to infer causal links between school resources and student outcomes based on observed variation in resources within regression models. Evidence in the United States has been extensively debated since Hanushek (1986, 1997) and Greenwald, Hedges, and Laine (1996), with few firm conclusions (Hanushek, 2003). For a comprehensive review for developing countries, see Glewwe, Hanushek, Humpage, and Ravina (2013).

This description of these methods is deliberately brief. Each category has an enormous body of supporting methodological research to ensure the design is applied correctly. This methodology is beyond the scope of this book, and readers should consult one of the many textbooks on identifying impacts in education and social science research (Angrist & Pischke, 2009; McEwan, 2015). Here, our focus is on highlighting the relevant attributes of each evaluation design for an analyst who is contemplating a cost or CE or BC analysis.

7.2.2. Identification Strategies With Cost Analysis

For an economist performing cost analysis, there are several factors to consider. The most obvious is that the appropriate identification strategy for cost analysis is the one that best identifies the effects of the intervention. If the method yields valid results for the outcomes of an intervention, then those results can be combined with cost information to perform CE or BC analysis.

For economic evaluation, there are strong reasons to prefer experimental methods. These reasons are in addition to the stronger "gold standard" claims of causality or internal validity that are typically associated with experimental methods.

First, with experimental research there is typically much more detail on the specifics of the treatment. These details should make it easier for the analyst to estimate the resources required to implement

[1] To provide one example, there is a long literature comparing the achievement of students who attend private schools to students who attend public schools (Coleman et al., 1966; Lubienski & Lubienski, 2013). Yet, families who send their children to private schools are often of higher SES, so it is necessary to control for this difference. Even with such controls, however, it is feasible that students in private schools are different in some important, but unobserved, ways. Perhaps their families place higher priority on education and so help their children learn at home in subtle and hard to observe ways. Using correlational analysis, the effects of private school may therefore be confounded with household resources.

the intervention. As Chapters 3 through 5 indicated, collecting cost information is far from straightforward in practice—not least because interventions and reforms are often very loosely specified. Having detailed information on the implementation of the intervention is therefore a sizable advantage for CE analysis. This advantage is particularly important when evaluating educational interventions that can be implemented flexibly. For example, literacy reforms such as Reading Recovery or Success for All have many components and levels. Programs such as Read 180 are implemented in diverse ways across sites. Programs such as Reading Partners involve several partner agencies. The actual resources used for these educational interventions are most easily calculated when they are delivered as part of an experimental research project.

A second advantage of the experimental method is that the analyst has direct, parallel information on the control group as well as the treatment group. For CE analysis, this parallel information is critical. The analyst may be able to estimate an effect size gain from a particular reform versus the status quo and may be able to estimate the net costs of the reform. However, it is often difficult to estimate the resources used by the comparison group. For example, an after-school intervention to help students complete high school may cost $4,000 per student to implement, and its effects may be precisely identified by a quasi-experimental study (e.g., where after-school enrollment is instrumented from exogenous variation in program availability in the local area). But it may not be obvious what resources the comparison group receives: Some students may be in other after-school programs, others may be in youth training programs, and others may be employed. This information is not typically collected—most likely because it is not available—when quasi-experimental and correlational methods are applied.

Information on the comparison group is critical for CE analysis (and all economic evaluations). As noted earlier, the recent cost analysis of Success for All establishes that the incremental cost of the program is very low; most schools already allocate resources to students for similar programs (Quint, Zhu, Balu, Rappaport, & DeLaurentis, 2015). The substantive distinction of Success for All is how—rather than how much—resources are allocated for the program. This distinction can best be illustrated with information on the treatment and control groups.

By contrast to the experimental method, it may be more difficult to perform CE analysis with results from quasi-experimental methods. The treatment may be well defined, but the counterfactual is typically not;

so the incremental cost of treatment relative to control is hard to calculate. This difficulty is also apparent for correlational studies, although these studies are further compromised if they have weak construct validity.

A third advantage of the experimental method relates to noncompliance and attrition. For cost analysis, it is important to distinguish students who are assigned to receive the treatment but do not (noncompliers) and to identify students who only partially comply with the treatment (attriters). We might expect that noncompliers do not receive any resources and so have zero costs. Therefore, the total cost for the treatment-on-the-treated (TOT) group should be lower than for the intent-to-treat (ITT) group. Similarly, those who only attrite from the treatment will have lower costs than those who complete the program (and presumably have costs that are closer to the control group). Experimental methods usually allow for a clear distinction of these two groups and hence for a more accurate estimate of the costs of the intervention.

These considerations are especially important when examining production function studies, which purport to show the relationship between inputs and educational outputs. Looking across the evidence for developing countries, Glewwe et al. (2013) identified 79 high-quality production function studies. The outcomes of these studies were test scores, and the inputs covered a wide array of potential school resource measures (some of these studies applied experimental methods). The results were summarized using the vote-count method, with each estimation result listed by sign and statistical significance (see Glewwe et al., 2013, Table 2.7). The evidence is strongly plausible for most resource indicators. More resource-intensive provision of books were clearly found to increase student learning, as was classroom furniture (e.g., desks), basic infrastructure (electricity, building structures and libraries), and basic classroom materials (e.g., blackboards). Similarly, more resources allocated to teachers—as reflected in their education levels, experience, and directly in their pay—were associated with increased learning (Glewwe et al., 2013, Tables 2.8 and 2.9). For the pupil-teacher ratio, results were less conclusive: Across 101 estimates, only 30 were statistically significant and in the expected direction (higher ratios impairing learning outcomes). But it may also be common in some countries to assign weaker students to smaller classes than those who are doing well or rural students to smaller classes because of the dearth of students in their catchment areas. Overall, these results indicate that, in line with theory and common sense, more resources do enhance learning outcomes.

Incorporating this evidence into a CE framework is not straightforward. First, the cost of each input (e.g., the desks or electricity) is unknown and may be very hard to estimate without a direct research inquiry. Second, even if the cost of the input is known, we cannot know whether the school was allocating resource to other inputs. A treatment school with the specific input (e.g., blackboards) may be sacrificing other inputs that are not fully measured (e.g., books) such that the resource levels are different rather than higher in the treatment school. Ultimately, these resource measures are not the same as costs in money terms. The most prominent example is the pupil-teacher ratio. One might think that a school with a low pupil-teacher ratio has relatively more resources. But it may be that the school has fewer management staff or less resource for libraries, for example. For production function studies that examine the effect of facilities on educational outcomes, we would need to amortize the value of improvements in facilities (Cellini, Ferreira, & Rothstein, 2010; Duflo, 2001). In general, we must be careful in taking results from production function studies, estimating costs, and then calculating CE ratios because of identification problems.

In summary, for economic evaluation it is especially important that the method used to identify impacts from a treatment can also be used to determine costs—both for the treatment and the counterfactual.

7.2.3. Evidence From Meta-Analysis

Researchers in the social sciences are increasingly using techniques of meta-analysis to arrive at estimates of effectiveness. Often, there are numerous—perhaps hundreds—of individual studies that explore the causal relationship between a particular educational alternative and an outcome such as achievement. Results from individual studies may vary considerably. It is difficult to extract meaningful conclusions from the overall body of findings without resorting to additional analytical techniques. Thus, many researchers use meta-analysis to estimate the "average" effect size of an alternative, which is typically used to support broad conclusions about its effectiveness (see Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, 2009).

Meta-analyses have been conducted in many areas of educational research, ranging from the effects of within-class ability grouping to class size reduction (see the review of interventions by Ahn, Ames, & Myers, 2012). Again, we leave aside direct methodology of meta-analysis (Valentine, Cooper, Patall, Tyson, & Robinson, 2010). Instead, we consider how meta-analysis might be used to derive an effectiveness measure for CE analysis.

From our perspective, the pertinent issue is whether meta-analytic summaries can be combined with costs data in order to provide a CE comparison of different educational alternatives. Instead of using a single estimate of effectiveness, is it necessary or preferable for CE analysis to use an estimate derived from a comprehensive meta-analysis?

The incorporation of meta-analytic results into CE analysis warrants a fair amount of caution (Levin, 1988, 1991). Meta-analysis provides an estimate of the average results from many different versions of a single class of interventions (e.g., computer-assisted instruction, ability grouping, or tutoring). However, CE analysis is fundamentally oriented toward providing concrete information to decisionmakers on whether specific programs or policies are desirable to implement. Instead of specifics, a meta-analysis can only provide a general judgment of whether a general variety of policy is effective "on average."

The problem becomes more severe when we attempt to incorporate costs. In prior chapters, we discussed the importance of clearly defining an alternative, providing a detailed account of the ingredients, and carefully estimating the cost of each ingredient. But the effect size from meta-analysis is based on a mixture of many different programs, precluding any conceptual or practical way to identify costs. The effect size does not refer to an implementable program alternative with a set of specific ingredients. Consider a hypothetical meta-analysis of adult tutoring programs in elementary schools. In practice, each of these programs might obtain its tutoring services in different ways. Some might pay on-duty teachers to spend time after school, whereas others might pay local adults the minimum wage to participate. Still others could receive voluntary tutoring services from parents. Faced by such heterogeneity of resource use, there is no obvious way to define the ingredients and costs of a single program.

Moreover, meta-analytic effectiveness estimates are reported after controlling for characteristics of the research evaluation. These characteristics are grouped into categories defined as units; treatment; observing operations; setting; and method (Ahn et al., 2012; Cooper, 2009). Some of these characteristics are almost certainly correlated with the costs of an intervention. For example, the units category may refer to the grades of the students and the scale of the intervention, the treatment category may include information on how long the intervention is implemented for, and the setting may include the locality of the intervention. Each of these domains will influence the costs of the intervention. Controlling for treatment duration is therefore, to some extent, controlling for costs.

Under stringent conditions, it may be acceptable to use meta-analytic results. Overall, meta-analytic results should not be incorporated in CE analyses unless the underlying situations are derived from replication trials of a single intervention and the meta-analytic outcome does not come from a model specification that controls for resources. If the specific studies all refer to different evaluations of precisely the same intervention, then it is more acceptable to ascribe a meaningful policy interpretation to the "average effect." When the intervention is precisely the same, it is more likely that the particular cost ingredients will be similar across studies. For example, a specific intervention—such as a "packaged" reading program—may use a prescribed amount of materials, physical space, time, and human resources, even if it is implemented and evaluated in many different contexts (the messy reality of program implementation, however, provides good reason to be skeptical that this proposition will always hold). Also, meta-analysis can be used as part of a sensitivity analysis to estimate upper and lower bounds for the effects of an intervention.

7.3. UTILITY ANALYSIS

Often a single measure of effectiveness does not fully describe a program's outcomes and is not a true reflection of the policymaker's preferences.

One technical solution, as noted previously, is simply to apply each effectiveness measure in a separate CE analysis. If all versions of the analyses yield the same or similar rankings, the multiplicity of outcomes is not salient. If, as is more likely, the analyses yield mixed results, the policymaker still has some information on which to base a resource allocation decision.

A more theoretically grounded solution is to derive a utility function. This function can then embody the relative value to the decisionmaker of increases in diverse educational outcomes. Utility is a shorthand way of describing the relative strength of preference or satisfaction that parents (or students or teachers) have for each outcome within a range of possibilities. It can be applied to measure of effectiveness. A utility function is intended as a map of the decisionmakers' preferences.

The tricky part is deriving a good estimate of the utility provided by each alternative—that is, to specify a utility function (mathematical representation) that incorporates all outcomes. This task is even more complex if the two outcomes are very different: if, for example, class size reduction improves only achievement in reading and is found to ameliorate externalizing behaviors. To help specify a utility function, researchers have developed techniques in "decision analysis," although these techniques have not been widely applied in education. This section reviews some of the most straightforward approaches. First, we will provide an overview of multiattribute utility theory, which serves as a convenient framework to organize the discussion. This is followed by a review of the methods for estimating utility and a discussion of whose preferences (or utility) should be measured.

Despite the clear relevance of utility theory (and also CU theory) to interventions with multiple outcomes, very few CU analyses have been performed across educational research (Ross, 2008). Of necessity, therefore, our discussion of utility focuses on methodological issues rather than examples of published studies. By contrast, CU analysis is widely practiced in health research, primarily because a consensus on how to measure utility is well established (Neumann, Thorat, Shi, Saret, & Cohen, 2015).

7.3.1. Multiattribute Utility Theory

Multiattribute utility theory is a complicated name for a fairly intuitive idea. An educational program produces outcomes in a multitude of categories: student achievement, student and teacher attitudes, and so on. Within each category, we could imagine a variety of subcategories. For example, student achievement can be divided into mathematics, reading, science, and so on. The literature on utility theory refers to each subcategory or measure of effectiveness as an "attribute." We shall adopt the latter term in the following discussion. Stakeholders may derive utility from—or have a preference for—each of these attributes. Multiattribute utility theory provides a set of techniques for accomplishing two tasks: (1) quantifying the utility derived from individual attributes and (2) combining the utility from each attribute to arrive at an overall measure of utility. The general tool for carrying out these tasks is called the multiattribute utility function.

Imagine that we exhaustively catalogued and evaluated the attributes of a particular educational program. We could use a simple notation to refer to each of these attributes: x_1, x_2, x_3, and so on, through the final attribute, x_m. These attributes are measured in their "natural" units. For example, gains on an achievement test might be expressed in percentage points, the number of test items, or months of learning gain. To perform CE analysis, we will need to express each attribute on a common "utility" scale. That is, we would like to describe the strength of preferences for a given increase in achievement, for an improvement in student attitudes, or for a change in any of the attributes.

We need to estimate a series of single-attribute utility functions: $\mathrm{U}_1(\mathrm{x}_1)$, $\mathrm{U}_2(x_2)$, and $U_3(x_3)$, through $U_m(x_m)$. The preceding notation is an efficient way of saying "the utility produced by the attribute x_1," "the utility produced by the attribute x_2," etc. In the next section, we will specify how to "convert" each attribute to a utility scale.

Once single-attribute utility functions are obtained, the next step is to combine them in an overall measure of utility. The tool for doing so is referred to as the multiattribute utility function. The overall utility from a given alternative (and its m attributes) is expressed as follows:

$$U(x_1, \ldots, x_m) = \sum_{i=1}^{m} w_i U_i(x_i)$$

It is nothing more than a weighted sum of the utilities produced by individual attributes. To make this more concrete, let us assume that the outcomes of a particular alternative are fully described by three attributes:

$$U(x_1, x_2, x_3) = w_1 U_1(x_1) + w_2 U_2(x_2) + w_3 U_3(x_3)$$

Prior to summing the three single-attribute utility functions, each is multiplied by an "importance weight" (w_1, w_2, and w_3). In general, the importance weights across all the attributes should sum to 1 (i.e., $\mathrm{w}_1 + \mathrm{w}_2 + \mathrm{w}_3 = 1$). Each weight should reflect the relative importance of each attribute to the stakeholders. For example, if $w_1 = 0.80$, $w_2 = 0.10$, $w_3 = 0.10$, then the overall utility of stakeholders is primarily determined by attribute x_1 with the other two attributes having lesser (and equal) importance. Below we specify how to elicit importance weights from stakeholders.

This type of multiattribute utility function is "additive": It involves simply adding up the weighted utilities of individual attributes. It makes intuitive sense to most people, and it can be usefully applied in a variety of circumstances. For example, if the attributes are reading, math, and science achievement, researchers might consider these as cognitive gains that can be summed. Nevertheless, the additive utility function is restrictive. It assumes that the preference for each attribute is independent of the preferences for the other attributes. This assumption may not be realistic: Families may care that children make moderate gains in all subjects rather than sizable gains in only one subject.

Before overall utility scores can be obtained, however, there are two remaining steps. First, we need to convert each attribute into a common utility scale that expresses the strength of preference for the attribute. That is, we need to define the functions—$U_1(x_1)$, $U_2(x_2)$, and so on—that describe exactly how additional units of the attributes are associated with utility. Second, we need to establish the weights—w_1, w_2, and so on—that reflect the relative importance of each attribute in overall utility. Toward accomplishing this, the following sections explore a few of the techniques that scholars in the field of decision analysis have devised.

7.3.2. Methods of Assessing Single-Attribute Utility Functions

This section describes several approaches to assessing single-attribute utility functions: proportional scoring, the direct method, and the variable probability method. To better illustrate each approach, we shall employ some hypothetical data on effectiveness. Imagine that we have just evaluated four separate programs for computer-assisted instruction of mathematics. The four alternatives (A, B, C, and D) are each evaluated according to a single attribute: mathematics scores. The test is composed of 25 items, results of which are presented in Table 7.2. In the following sections, we will convert these attribute scores to a utility scale.

Proportional Scoring

The first method, proportional scoring, is simply a linear rescaling of each attribute to a common utility scale. The rescaling can be

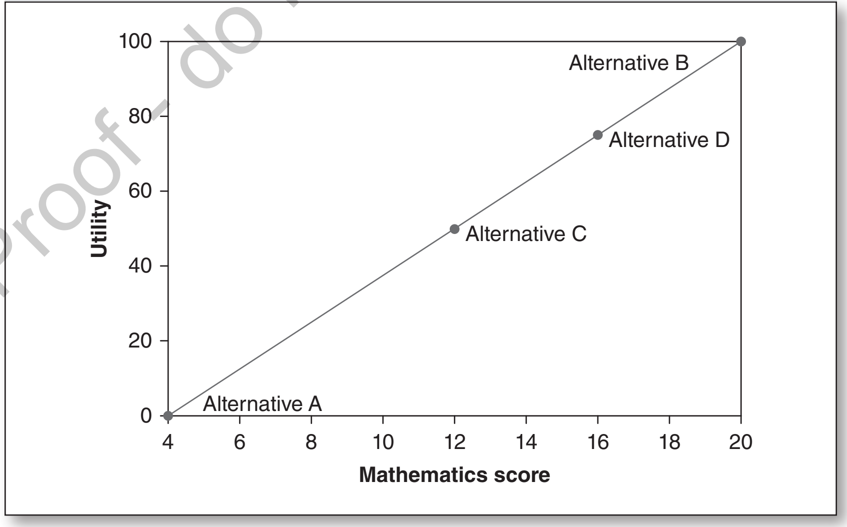
Table 7.2 Hypothetical Data From an Evaluation of Four Programs for Computer-Assisted Math Instruction

	Mathematics Scores
Alternative A	4
Alternative B	20
Alternative C	12
Alternative D	16

accomplished via graphical or mathematical means. In Figure 7.1, we provide a graphical representation of proportional scoring. Each mathematics score is plotted on the *x* axis, ranging from the value of the lowest-scoring alternative to that of the highest-scoring alternative. The utility scale, on the *y* axis, ranges from 0 to 100. The low and high values of the utility scale are arbitrary—we could just as easily set the end points at any values. The same utility scale must be shared by each of the attributes that we assess (and eventually combine into a single measure of utility).

As shown in Figure 7.1, the lowest score on mathematics is assigned a utility of 0 and the highest a utility of 100. The straight lines connecting these points imply that increasing mathematics scores lead to constant increases in utility (in this case, a 4-point increase in mathematics scores produces a 25-point increase in utility). Of course, this is an assumption that we are making. We have no direct evidence that people really evince this preference structure. It might be that when reading scores are low, a small increase leads to a substantial utility increase, but when they are higher, the same increase in scores leads to a somewhat smaller gain in utility. This would be represented by a curvilinear, rather than linear, utility function. Later on, we will allow for this possibility.

Figure 7.1 Assessing Utility Functions With Proportional Scoring



We could derive the same utility scores mathematically, without resorting to graphs. The formula is quite simple:

$$U(x) = \frac{x - \text{Lowest}}{\text{Highest} - \text{Lowest}} \times 100$$

Applying the formula for a reading score of 12 (Alternative C) yields a utility score of 50 (this can be verified by examining Figure 7.1):

$$U(12) = \frac{12 - 4}{20 - 4} \times 100 = 50$$

In a sense, proportional scoring isn't really a "method" because it does not rely on the expressed preferences of stakeholders. It simply assumes that increasing amounts of an attribute are linearly (proportionally) associated with utility.

The Direct Method

Instead of using proportional scoring, we could obtain direct input from individual stakeholders on the utility that they derive from varying amounts of an attribute. The simplest approach for doing so is the direct method. To apply the direct method, one identifies the low and high values on the relevant attribute scale. In this case, the low mathematics score is 4 and the high score is 20. As before, these are arbitrarily assigned low (0) and high (100) values, respectively, on the utility scale. The respondent is then asked to directly rate the preference for middle levels of the attribute, relative to these end points. In our example, the middle levels are the mathematics scores that were obtained by the middle alternatives. For comparison's sake, it would also be helpful to rate other possible scores. We could administer a survey to education professionals or parents asking them to rate scores on a mathematics achievement test. Assume such a process turned up the following results:

$U(4) = 0$ (arbitrary assignment)

$U(8) = 40$ (judgment, relative to arbitrary assignment)

$U(12) = 75$ (judgment, relative to arbitrary assignment)

$U(16) = 95$ (judgment, relative to arbitrary assignment)

$U(20) = 100$ (arbitrary assignment)

The mathematics scores and corresponding utilities are plotted in Figure 7.2. A researcher could use visual means to draw a smooth curve through the points. Alternatively, many researchers use statistical methods to find the curve that provides the best "fit" to the data. In this case, the data suggest a curvilinear relationship between mathematics scores and utility. More specifically, increasing mathematics scores tend to increase utility, but at a decreasing rate. Of course, utility functions can assume many different shapes depending on the survey responses. (The structure of the prior example was borrowed from von Winterfeldt and Edwards (1986); see also Gray, Clarke, Wolstenholme, and Wordsworth [2011].)

The Variable Probability Method

The variable probability method also calls upon stakeholders to assess their preferences for varying amounts of a given attribute. However, it requires a different sort of thought experiment than the direct method. Imagine that you are able to choose between two different options. On the one hand, you could opt for a gamble in which the "winning" hand leads to the highest attribute score (in this case, a mathematics score of 20) and the "losing" hand produces the lowest (a mathematics score of 4). The probabilities of attaining the highest and lowest scores are, respectively, *p* and (1-*p*). Instead of this risky option,

Figure 7.2 Assessing Utility Functions With the Direct Method

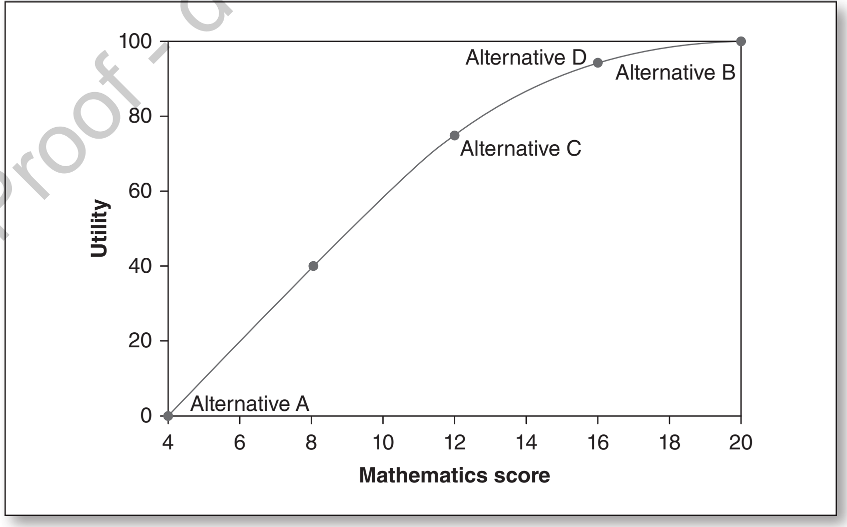
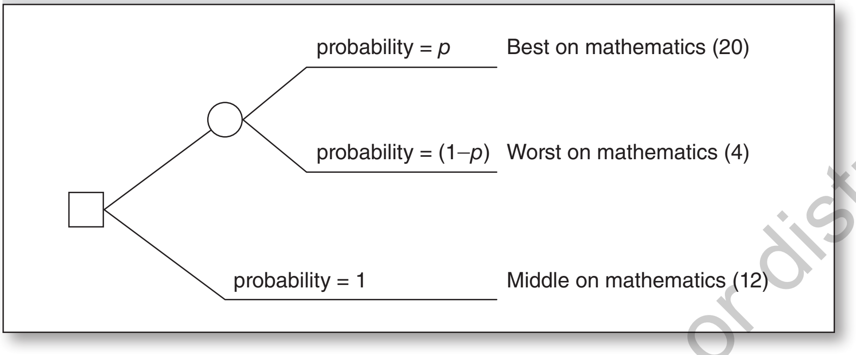


Figure 7.3 Assessing Utility Functions With the Variable Probability Method



you could obtain a given mathematics score with certainty. For the time being, let's fix this middle score at 12. This particular gamble is represented by the decision tree in Figure 7.3.

To assess the utility of the middle score, individuals choose the probability (p) that makes them indifferent between the risky alternative (with a potentially high or low payoff) or the riskless alternative (with a middling payoff). Let us say, for example, that we suggested an initial probability of 0.99. That is, individuals would be faced with the option of receiving the best score with a probability of 0.99 (and, conversely, the worst score with a probability of 0.01) or receiving a middling score with certainty. Many individuals would likely find the risky option to be most attractive.

What if we suggested an initial probability of 0.01 instead of 0.99? In this case, chances are that most individuals would not favor a gamble that offered such a small probability of an attractive payoff. Instead, they might prefer the certainty of obtaining a middle score.

Between 0.99 and 0.01, there is a probability at which individuals would be indifferent between the two options. In the case of Figure 7.3, suppose that a probability of 0.60 leads to indifference for a particular individual. We can then interpret this probability as the utility of a mathematics score of 12 (with the endpoints of the utility scale set at 0 and 1). In order to employ the same utility scale as prior examples, we multiply 0.60 by 100, yielding a utility of 60. The same exercise is repeated for several different mathematics scores. Doing so produces a number of pairs of mathematics scores and their associated utilities. These can be graphed, just as we did in Figures 7.1 and 7.2.

7.3.3. Methods of Assessing Importance Weights

After single-attribute utility functions are defined for each attribute, we require some method for obtaining the relative weight or "importance" of each attribute in overall utility. The two general approaches are the direct method and the variable probability method.

The simplest version of the direct method asks individuals to "allocate" a total of 100 points among attributes, according to their relative importance. Let's say that mathematics scores are considered by individuals to account for about half of overall utility and, consequently, are assigned 50 out of 100 points. Computer literacy is the next most important attribute and is assigned 30 points. Lastly, student satisfaction receives 20 points. Each estimate is divided by 100 in order to obtain a set of three importance weights—0.50, 0.30, and 0.20—that sum to 1. In other variants of the direct method individuals are asked to rank attributes in order of importance.

With the variable probability method, individuals are asked to choose between two options. One is a gamble with two possible outcomes (e.g., the best test score on all attributes with probability p or the worst test score on all attributes with probability $1-p$). The other option is a certain outcome (e.g., the best test score on just one test). If the probability (p) is 0.99, many individuals would choose the gamble; if the probability is 0.01, the gamble is much less appealing. Between these two values of p, there lies a probability that would cause an individual to be indifferent between the two options. This probability can be interpreted as the importance weight for a particular test. Once importance weights are estimated for all attributes, they should sum to unity.[2]

Thus far, we have addressed several techniques for assessing the utility of individuals. However, we need to specify exactly *whose* preferences should be assessed. There are at least three groups that might be considered: (1) the entire population in a given community; (2) the population that is directly affected by an intervention

[2] If they are close (but do not sum to 1) we can normalize them by dividing each individual weight by the sum of the weights. If the sum is not close to 1, this is a signal that the additive utility function does not adequately represent an individual's preferences (Clemen, 1996). The analyst might need to use more complex versions of the utility function that incorporate interactions among the attributes. These were briefly mentioned in a previous section. For further details, the reader is encouraged to consult Clemen (1996) or Keeney and Raiffa (1993).

(such as families with children enrolled in school); and (3) a smaller group of representatives such as teachers, administrators, or school board members. In choosing among these, evaluators should also consider how the results from CU analysis might improve decision-making or change practices. If, for example, the research is motivated to help parents choose between reading strategies, then parental utility is the relevant preference. In keeping with our general approach, the presumption is that CU analysis should attempt to measure the preferences of an entire community.

Preferences can be elicited through survey responses.[3] However, sampling should be performed carefully because it may be hard to judge the degree of variability in preferences. In education research, there is simply little evidence to guide us. Earlier educational CU analysis only elicited the preferences of small groups of administrators or other stakeholders (e.g., Fletcher, Hawley, & Piele, 1990; Lewis, Johnson, Erickson, & Bruininks, 1994). More recently, Ross (2008) found that different groups of professionals had very different weightings and ratings of library services within a local school district. Finally, in a study of reading outcomes, Simon (2011) found that the preferences of reading professionals varied depending on whether the students were average readers or struggling readers. Specifically, these professionals gave greater weight to phonemic awareness outcomes for struggling versus average readers and lesser weight to fluency (Simon, 2011, Table 27). If the preferences vary significantly across groups, the utility may not be valid.

7.3.4. Using Utility Measures

Few educational interventions use utility measures as measures of effectiveness (e.g., Ross, 2008; Simon, 2011). A full illustration is given in Example 7.1, where the outcome is special education programs.

[3] In some cases, however, it will not be possible to obtain the views of a large sample of community members. Perhaps time is a binding constraint, or the monetary costs of a community survey are judged to be prohibitive. There are two alternatives that might be pursued. First, one can assess the preferences of a representative sample of parents or students who are directly affected by the intervention. Second, one can obtain the views of appropriate representatives of the community, such as school board members or elected officials of civic and community organizations. In other cases, it may be possible for administrators or teachers to determine the utility of the alternatives.

Example 7.1 Cost-Utility Analysis of Special Education Alternatives (Part 1)

The outcomes of special education programs are difficult, if not impossible, to express with a single measure of effectiveness (or attribute). As such, multiattribute utility theory seems especially appropriate. Here, we describe the utility step; in Chapter 8, we link these utility measures to costs to perform cost-utility (CU) analysis.

Darrell Lewis and his colleagues (1994) set out to compare the utility (and costs) produced by three different administrative structures for special education. These alternatives were (a) an independent school district (offering special education services to all students within the locality), (b) an intermediate school district (jointly offering services for students primarily with low-incidence disabilities), and (c) a joint powers special education cooperative (with districts sharing delivery of special education). At issue was which administrative structure would yield the highest utility and at the lowest cost.

The first step is therefore to measure utility. In collaboration with a group of stakeholders—including teachers, administrators, and parents—the evaluators defined the attributes by which the success of alternatives would be judged. These attributes are itemized in the first column of the following table. These are grouped into four categories: (1) student participation in school life, (2) satisfaction with the program, (3) program accomplishments, and (4) program processes.

Estimating the Utility of Special Education Alternatives

	Independent District Alternative				
Interventions	**Importance Weight**		**Unweighted Attribute Utility (0–100)**		**Weighted Attribute Utility**
Student participation in school life					
Access to educational/social experiences	0.09	×	32.5	=	2.9
Participate in extracurricular/ social activities	0.07	×	13.3	=	0.9
Participate in mainstream programming	0.09	×	80.0	=	7.2
Satisfaction with program					
Parents express satisfaction	0.05	×	84.7	=	4.2
Students express satisfaction	0.05	×	48.0	=	2.4
Teachers and administrators express satisfaction	0.04	×	82.7	=	3.3
Public expresses satisfaction	0.05	×	90.0	=	4.5

(Continued)

(Continued)

	Independent District Alternative				
Interventions	Importance Weight	Unweighted Attribute Utility (0–100)		Weighted Attribute Utility	
Accomplishments of program completers					
Demonstrate appropriate social behaviors	0.06	×	77.5	=	4.7
Live in independent/ semi-independent settings	0.06	×	54.0	=	3.2
Have social and recreational networks	0.06	×	89.9	=	5.4
Participate in meaningful vocational settings	0.06	×	88.5	=	5.3
Complete all years of offered schooling	0.04	×	100.0	=	4.0
Process of program					
Provides appropriate curriculum components	0.10	×	74.7	=	7.5
Provides training and support for parents	0.08	×	59.3	=	4.7
Provides appropriate staff support	0.09	×	40.0	=	3.6
Sum	1.00				63.9

Source: Adapted from Lewis et al. (1994, Tables 3 and 6).

The same group of stakeholders assigned importance weights to each attribute using the direct method. Individuals ranked all the attributes in order of their importance, with the most important being assigned a value of 100. The rest of the attributes were assigned lesser values, relative to 100, and all these values were normalized to sum to 1. The final importance weights are presented in the second column of the table.

The evaluators then visited school districts and conducted surveys to collect the performance data on each attribute. These attributes were measured on a variety of scales. However, it was necessary to convert each of these to a common utility scale, with the lowest possibility utility of each attribute specified as zero and the highest utility as 100. To convert each attribute score, the evaluators used the proportional scoring method. The third column in the table presents the unweighted attribute utilities for one of the three alternatives—independent districts.

The final step is to combine importance weights and unweighted utilities in order to arrive at an overall measure of each alternative's utility. To do so, the evaluators employed the additive multiattribute utility function. Each attribute's utility was multiplied by its respective importance weight (see the fourth column). The weighted utilities were then summed, thereby yielding the overall utility of the alternative. The table shows that the overall utility of the independent district alternative is 63.9. Other calculations, not shown in the table, implied a utility of 70.4 for the intermediate alternative, and 65.2 for the cooperative alternative.

The results suggest that the intermediate alternative is the most attractive: It provides the highest level of utility. However, it is important to combine these results with cost estimates in order to determine which alternative provides a given level of utility at least cost. We report this CU analysis in Chapter 8.

Source: Adapted from Lewis et al. (1994).

One very common, well-accepted utility measure is the quality-adjusted life year (QALY). The QALY takes the value 1 for a year of life in perfect health and is adjusted downward to zero for progressively worse health conditions. Health interventions are frequently evaluated according to their effects on life expectancy. That is, by how many years does a particular medical treatment tend to lengthen one's life? While a useful means of evaluating some interventions, life expectancy still does not capture the quality of life or the satisfaction that individuals may derive from additional years of life. Two medical treatments may each add 2 years to an individual's life. Yet, if one of these leaves the individual significantly impaired or incapacitated, then it is clearly less desirable. To estimate the QALYs that are produced by a medical treatment, it is necessary to estimate quality-of-life weights that reflect the satisfaction derived from different health states. These weights can be obtained using a range of methods—for example, the standard gamble or time trade-off method (for a review of these, see Weinstein et al., 2009; Whitehead & Ali, 2010). Increasingly, given its acceptance in health sciences, researchers are using QALYs as a way to value educational interventions. For example, Muennig, Fiscella, Tancredi, and Franks (2010) estimate that a high school graduate will accumulate an additional 2.4 QALYs over their lifetime compared to a dropout. Schoeni, Robert, Dow, Miller, and Pamuk (2011) estimated a range of incremental QALYs by education level, also finding large QALY gains for high school graduates over high school dropouts. Educational interventions with a specific focus on child health may therefore rely on an established utility measure.

Finally, despite the lack of explicit utility measures available for education researchers, it is worth noting that many effectiveness measures are based on opinions. For example, college ratings are a mathematical combination of attributes where the weightings are based on survey information, student engagement indices are derived from opinion-based responses of students, and teacher and faculty competence are often based on student evaluations (see respectively Pike, 2004; Spooren, Brockx, & Mortelmans, 2013; Webster, 2001). These measures are artifacts such that they should be justified based on how accurately they reflect the preferences of decisionmakers. In cases where there are no obvious utility measures of effectiveness, analysts might need to perform their own survey as to which outcomes are most valuable and how each outcome should be weighted.

7.4. CONCLUSIONS

In this chapter, we have reviewed what makes for a good effectiveness measure for the purposes of CE. The use of impact evaluations for BC analysis is similar but involves a very particular step—turning the effect into a money value. We address this in Chapter 9.

We do not wish to understate the challenges involved in choosing a proper measure of effectiveness and the dangers involved in using a poor measure. It makes little sense to invest time and resources in accurate cost measurements and a rigorous evaluation design if the measure of effectiveness is not suitable. That said, much of our discussion is about what makes for a good effectiveness measure per se, which is—or should be—the focus of all impact evaluations. Of necessity, the effectiveness measure should fully reflect the objectives of the intervention, and it should be expressed as a single number (even as that number may be a composite of several constructs or derived from a utility function). Preferably, the effectiveness measure should be easy to interpret on a continuous scale.

It is also preferable that the measure is estimated using an experimental method. But we note that this preference is not because experimental methods are more reliable and internally valid. Rather, it is because the experimental method allows us to collect much more information on how the intervention is implemented and therefore how much it costs as well as equivalent information for the counterfactual.

With a valid, reliable, and meaningful effectiveness measure, we can combine this with information on costs to calculate the CE ratio. This is the subject of the next chapter.

Discussion Questions

1. What criteria should be applied when choosing an effectiveness measure that is to be applied in CE analysis?
2. What are some potential threats to the validity of a measure of a program's effectiveness? What estimation and measurement methods can be used to overcome these threats, and what are some limitations of these methods?
3. What is meta-analysis? What are some reasons why its applicability to CE analysis might be limited?

Exercises

1. As an analyst for a school district, you review the evidence on programs to increase attainment in high school. You identify the following studies:

Program	Treatment Group Size	Percentage Point Gain Over Control Group in High School Graduation Rate
Talent Search	3,930	10.8
Job Corps	3,940	17.0
JOBSTART	1,028	15.1
New Chance	1,240	9.2
National Guard Youth ChalleNGe Program (NGYCP)	596	19.8

How would you express these results for CE analysis? Which program do you recommend? What other information on effects might be useful?

4. You have been asked to perform CE analysis on a series of middle school math programs, each of which has undergone an experimental evaluation comparing its effects to those of the standard math curriculum on a series of assessments. The following table summarizes the results:

Program	Sample	Measure	Result
Alpha Math	Two classes of sixth-grade math students	Effect size gains on state math assessment	0.1**
Acing Algebra	Four groups of eighth graders in remedial math classes (two years below grade level)	Effect size gains on a program-specific assessment	0.25***
Sigma!	Three groups of sixth-grade math students (performing at grade level)	Effect size gains on standardized math test	0.15*
Primed for Algebra	Three groups of seventh-grade students (one year below grade level)	Effect size gain on a standardized math test	0.08

Note: *p <= .05, **p<= .01, ***p<= .001

Which programs would you compare to one another in CE analysis? Which ones would you recommend against comparing in a CE framework? Why? What other factors would you consider in making comparisons?

3. In an experimental test of a financial incentive program for community college students, Barrow, Richburg-Hayes, Rouse, and Brock (2014) estimated the following results (all statistically significant):

	Baseline after two semesters	Program effects
Enrolled in any course (%)	49.6	15.0
Total credits attempted	4.9	1.2
College-level credits earned	2.1	0.9
Total credits earned	2.8	1.1

Which measure is most appropriate for CE analysis?