IDB
Development Effectiveness

Impact-Evaluation Guidelines
Technical Notes
No. IDB-TN-332
December 2011

Cost-Effectiveness Analysis of Education and Health Interventions in Developing Countries

Patrick J. McEwan

Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank

Cost-Effectiveness Analysis of Education and Health Interventions in Developing Countries

Impact-Evaluation Guidelines

Patrick J. McEwan

Inter-American Development Bank

2011

Patrick J. McEwan. Associate Professor of Economics, Wellesley College. pmcewan@wellesley.edu

Cost-Effectiveness Analysis of Education and Health Interventions in Developing Countries

Abstract[1]

Patrick J. McEwan[2]

High-quality impact evaluations, including randomized experiments, are increasingly popular, but cannot always inform resource allocation decisions unless the costs of interventions are considered alongside their effects. Cost-effectiveness analysis is a straightforward but under-utilized tool for determining which, of two or more interventions provides a (non-pecuniary) unit of effect at least cost. This paper reviews the framework and methods of cost-effectiveness analysis, emphasizing education and health interventions, and discusses how the methods are applied in the literature.

JEL Classification: H43, I25, Z18

Keywords: Cost-Effectiveness; Cost-Benefit; Impact Evaluation.

[1] I am grateful to Henry Levin for prior collaborations that greatly informed my thinking. Paul Winters and Francisco Mejia provided helpful comments on an earlier draft.Associate Professor of Economics, Wellesley College. pmcewan@wellesley.edu

[2] Associate Professor of Economics, Wellesley College. pmcewan@wellesley.edu

Table of Contents

1. Introduction

In 2000, the United Nations Millennium Declaration established ambitious goals for poverty reduction, focusing on education and health outcomes. But the route to achieving such goals was not clear: there were thousands of competing interventions to reduce poverty, and a research base not always capable of identifying the most cost-effective options (Duflo and Kremer, 2005; Duflo, 2004; Savedoff, Levine, and Birdsall, 2006). Fortunately, the quantity of impact evaluations in education and health grew rapidly, and they increasingly applied high-quality experimental research designs.[3] The growing number of impact evaluations has facilitated reviews of the most effective interventions in education and health.[4]

Such reviews provide advice about how to allocate scarce resources across a range of competing interventions, partly by ruling out interventions with zero or even harmful effects. But, as authors note, it is difficult to choose among a range of effective interventions unless impacts are considered alongside costs. Consider the popular education intervention of reducing the number of students per classroom in primary schools. Research in the United States and Latin America has found that class size reduction is an effective way of increasing students' test scores, and that its effects on test scores may even be larger than alternate interventions (Schanzenbach, 2007; Urquiola, 2006). However, class size reduction may still be less *cost-effective*, since it costs more than competing interventions to raise test scores by an equivalent amount (Levin et al., 1987; Loeb and McEwan, 2010).

A substantial literature codifies methods to measure and compare the cost-effectiveness of education and health interventions, but it has often focused on developed countries (Levin and McEwan, 2001; Drummond et al., 2005; Gold et al., 1996). In developing countries, the cost-effectiveness literature has mainly grown in health policy—see especially Jamison et al. (2006a, 2006b)—but recent experiments in education have often been accompanied by cost analysis. This paper provides an updated review of methods and applications of cost-effectiveness

[3] Prominent examples included the experimental evaluations of conditional cash transfers in Mexico and school-based deworming treatments in Kenya (Schultz, 2004; Skoufias, 2005; Miguel and Kremer, 2004).

[4] See, for example, Rawlings and Rubio (2006) and Fiszbein and Schady (2009) on evaluations of conditional cash transfers; Holla and Kremer (2009) on a broader class of education and health interventions that modifies prices; and Zwane and Kremer (2007) on the effectiveness of interventions to reduce diarrheal diseases in developing countries. Bouillon and Tejerina (2006) summarize impact evaluations from broad array of evaluation literature in health, education, and other sectors.

analysis, with an emphasis on education and health interventions in developing countries.[5]

Section 2 describes a general framework for cost analysis. In doing so, it highlights the key challenge of conducting cost-effectiveness analysis (CEA): to credibly identify the incremental costs and incremental effects of two or more interventions. The intervention(s) with relatively lower incremental costs per unit of incremental effect are better candidates for investment. Although it is not the main topic of the paper, section 3 briefly reviews methods for measuring effects, focusing on high-quality experimental and quasi-experimental designs.[6] Section 4 describes an intuitive approach to estimating costs—the ingredients methods—that it practiced in similar forms across many fields. It relies on the exhaustive specification of an intervention's cost ingredients and their prices, and judicious cost comparisons that adjust for price levels, time preference, and currency (Levin and McEwan, 2001). Section 5 provides a summary of cost-effectiveness analysis as it currently being conducted in developing countries, ranging from single-study CEA embedded within a randomized experiment to ambitious attempts to construct CEA "league tables" of multiple interventions. Section 6 considers common issues in conducting a CEA, including sensitivity analysis, external validity of cost-effectiveness ratios, and the appropriate steps to follow in conducting ex post and ex ante CEA. Section 7 concludes.

[5] Dhaliwal et al. (2011) conduct a similar review and reiterate many themes of this paper.

[6] For more detailed reviews, see DiNardo and Lee (2010), Imbens and Wooldridge (2009), Murnane and Willett (2011), and McEwan (2008).

2. Framework and Definitions

2.1 Types of Cost Analysis

Cost analysis falls into two broad categories: cost-benefit analysis (CBA) and cost-effectiveness analysis (CEA). A third approach, cost-utility analysis (CUA), is often implemented as an extension of CEA. All methods presuppose a well-specified intervention and a no-intervention condition, or control group, against which the intervention is compared. In general terms, an intervention uses human, physical, or financial inputs to improve individuals' education, health, or labor market outcomes. The intervention may be a small-scale program (school-based distribution of deworming drugs or textbooks) or a large-scale policy shift (nationwide elimination of school fees in public schools).

The costs of an intervention, *C*, are the opportunity costs of resources used in the intervention versus the no-invention control. Section 4 describes cost analysis in more detail, but two issues merit emphasis. First, *C* only reflects the cost of *additional* resources used in the intervention. Indeed, health economists engaged in CEA refer to *C* as the incremental costs of an intervention, and this paper adopts that term. Second, costs include any resource with an opportunity cost, even "free" resources such as volunteer labor. Such resources have an opportunity cost because they require the worker to forgo another valuable opportunity, and are costly to society.

2.2. Cost-Benefit Analysis

The fundamental difference between CBA and CEA lies in the measurement of the incremental outcomes of an intervention as (1) incremental benefits or (2) incremental effects. In CBA, the incremental benefits of an intervention are the monetary gains in social surplus created by the intervention (see Boardman et al., 2011 for a theoretical discussion). In practical terms, CBA of investments in human capital usually measure benefits as the additional earnings and tax revenues received by participants and governments, respectively. In other circumstances, benefits may be measured as averted costs: that is, monetary costs to society averted as a result of the intervention, such as reduced crime. Sometimes incremental benefits can be directly estimated in long-run experimental or quasi-experimental impact evaluations, but it is more common that benefits are projected and estimated based on shorter-term evaluations (see section 3 for additional discussion).

Once incremental benefits, *B*, are calculated, the value $B-C$ represents the net benefits

of a single intervention. More accurately, since benefits and costs are often distributed unevenly throughout time, we need to estimate the Net Present Value (NPV) of an intervention. The NPV is $\sum_{t=0}^{n}\frac{B_t}{(1+r)^t}-\sum_{t=0}^{n}\frac{C_t}{(1+r)^t}$, where incremental benefits (*B*) or costs (*C*) may be received or incurred immediately, at $t = 0$, or up to n years in the future. A lower discount rate r reflects a more "future-oriented" decision-maker that discounts future benefits (or costs) less heavily.[7]

The NPV has a simple and convenient interpretation as the absolute desirability of the intervention (whether positive or negative). Its magnitude can also be compared to NPVs of other interventions, both within a given sector (education or health) and across disparate and competing sectors (infrastructure).

2.3. Cost-Effectiveness Analysis

In CEA, incremental effects are expressed in non-monetary units. In education, the effects may include quantity measures such as school enrollment, attendance, completion, or overall years or degrees attained; and quality measures such as cognitive development, academic achievement, or non-cognitive skills. In health, the outcomes may include clinic enrollment or attendance, health incidents averted (e.g., respiratory or diarrheal illness), life years saved, or improved quality-of-life. Presuming that the incremental effect of an intervention, *E*, can be credibly identified—an issue revisited in section 3—the incremental cost-effectiveness ratio (CER) is $\frac{C}{E}$. It represents the incremental cost per unit of incremental effect (i.e., the cost of enrolling another child in school, or the cost of saving an additional life). Authors also report effectiveness-cost ratios ($\frac{E}{C}$, or units of incremental effect per unit of incremental cost), although it is less common.

In practice, *E* is often taken directly from an impact evaluation in which effects are expressed as an average treatment effect in a sample of individuals. For example, an intervention may be found to increase a child's expected probability of enrolling in school by 0.04, or to increase the expected test score of a student by 0.2 standard deviations. In such cases, *C* is expressed in expressed in similar terms, as the incremental cost per student subjected to the

[7] Related measures include the internal rate of return (the "break-even" value of *r* that equalizes discounted benefits and discounted costs), and the benefit-cost ratio (discounted benefits divided by discounted costs). The NPV is typically preferred for reasons outlined in Boardman et al. (2011), although education economists frequently report the rate of return.

intervention.

Table 1 summarizes a cost-effectiveness analysis of Kenyan education interventions to improve child test scores, adapted from Kremer et al. (2004). The authors conducted an experimental impact evaluation of a program that provided merit scholarships for adolescent girls who scored well on exams. The average treatment effect was 0.12 standard deviations (a common metric for expressing test score gains). The incremental cost per pupil was $1.69, implying a CER of $1.41 per 0.1 standard deviations.[8] Unlike the net present value in a CBA, the CER of a single intervention cannot be used to judge its absolute desirability, because there is no means of weighing pecuniary costs against non-pecuniary effects. The CER can be compared to those of other interventions, presuming that effects are measured in the same units. Among several effective interventions, which incurs lower costs to increase test scores by a given amount?

Table 1: Cost-Effectiveness Ratios of Education Interventions in Kenya

	Effect	Cost	CER	CER
Intervention	Average test score gain	Cost per pupil (excluding transfers)	Cost per pupil per 0.1 gain (excluding transfers)	Cost per pupil per 0.1 gain (including transfers)
Girls scholarship program				
Busia and Teso districts	0.12	$1.69	$1.41	$4.94
Busia district	0.19	$1.35	$0.71	$2.48
Teacher incentives	0.07	$0.95	$1.36	$4.77
Textbook provision	0.04	$2.24	$5.61	$5.61
Deworming project	≈0	$1.46	--	--
Flip chart provision	≈0	$1.25	--	--
Child sponsorship program	≈0	$7.94	--	--

Source: Adapted from Kremer et al. (2004, p. 52). Results for teacher incentives, textbook provision, deworming, flip charts, and child sponsorship are from, respectively, Glewwe et al. (2003); Glewwe et al. (1997); Miguel and Kremer (2004), Glewwe et al. (2004); and Kremer et al. (2003).

In Table 1, the authors calculated CERs for other interventions, using other Kenyan experimental evaluations, including a teacher incentive program, textbooks and flip chart provision, and school-based deworming. The effect of some interventions could not be

[8] The cost estimate excludes transfer payments (scholarships), since the cost to the government agency is exactly offset by the benefits to students. However, the cost does account for the deadweight loss from raising tax revenues for transfer payments.

statistically distinguished from zero in the impact evaluation, implying an infinite CER, and removing them from consideration (indeed, one of interventions judged to have zero effect was also the most costly). The CERs suggest that scholarships and teacher incentives are similarly cost-effective ($1.41 and $1.36 per 0.1 standard deviations, respectively), and much more so than textbook provision ($5.61 per 0.1 standard deviations).

The CERs, like any estimate of program impact, do not provide unambiguous guidance about resource allocation. For example, Table 1 indicates that the scholarship program as implemented in the Busia district is relatively more effective and cost-effective than the larger program. It also happens to be more cost-effective than teacher incentives. The result highlights (1) the importance of assessing the sensitivity of CERs to alternate assumptions about costs and effects; and (2) carefully considering the external validity of estimated CERs, especially when generalizing results beyond the population and setting of the original evaluation. Each issue is discussed further in section 6.

The Kenyan example reveals an inherent challenge of a CEA that is not present in a CBA. Most social interventions pursue multiple objectives. It is possible that an intervention is the most cost-effective option for increasing one outcome, but not another. The deworming project in Table 1 was among the least cost-effective options for raising test scores because it had a zero effect. However, it was very effective at raising school participation and also among the most cost-effective options for doing so (Dhaliwal et al., 2011; Miguel and Kremer, 2004). As another example, conditional cash transfer policies have multiple goals: reducing short-run poverty, increasing the quantity and quality of child and mothers' health, and increasing the quantity and quality of education received by children and adolescents.

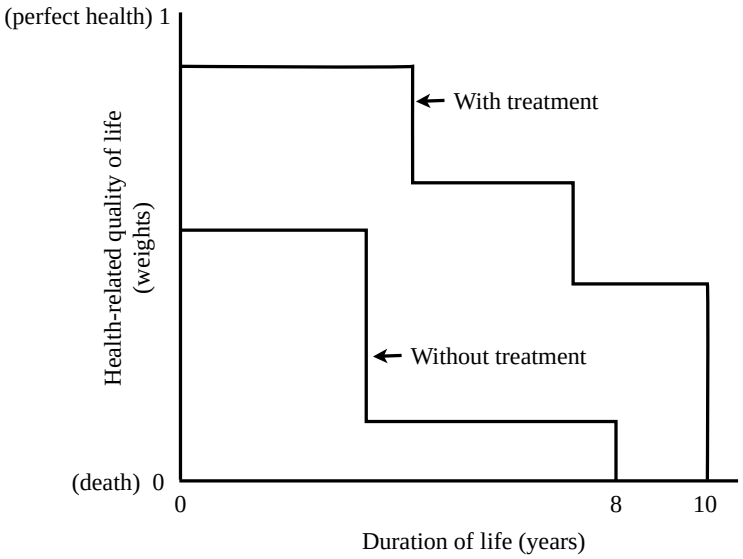
In such cases, most authors present CERs for each outcome and note important differences in rankings. Some authors further attempt to conduct a full CBA by estimating and aggregating the monetary benefits of two or more measures of outcomes, often using additional analysis of secondary data and assumptions (see section 3). A third option is to incorporate measures of "utility," where utility connotes satisfaction and does not always bear a close resemblance to theoretical concepts from microeconomics.

2.4 Cost-Utility Analysis

A cost-utility ratio $\frac{C}{U}$ reflects the incremental cost per unit of incremental utility. CUA is most common in health, where interventions often have the dual objectives of increasing life expectancy and also improving the quality of each year lived. Some interventions succeed in extending the number of years lived in poor health, while others improve general health without extending life. To compare the cost-effectiveness of these interventions, health economists calculate the incremental Quality-Adjusted Life Years (QALYs) produced by a health intervention, which is the denominator in a cost-utility ratio.

The idea of QALYs is illustrated in Figure 1. Imagine that one evaluates a medical treatment (relative to a no-treatment condition) and determines that it extends life expectancy by 2 years (from 8 to 10), measured on the x-axis. It also improves quality of each year lived, measured on the y-axis, where 1 indicates perfect health, and 0 indicates death, and intermediate values indicate degrees of impairment. The gain in QALYs that is produced by the treatment is calculated as the area between the two descending lines. Since incremental QALYs are unevenly distributed across years, it is standard to discount QALYs using the same formula and discount rate *r* used to discount monetary costs unevenly distributed across time (Gold et al., 1996).

Figure 1: An Illustration of Quality-Adjusted Life Years (QALYs)



Source: Adapted from Drummond et al. (2005).

The estimation of QALYs requires quality-of-life weights that reflect satisfaction derived from different health states. An extensive literature in health economics describes methods that usually involve surveying a sample of individuals and eliciting subjective estimates (e.g. Drummond et al. 2005; Muennig, 2008). In developing countries, it is more common to estimate and report Disability-Adjusted Life Years (DALYs), as in the well-known World Bank Development Report (World Bank, 1993) and its update (Jamison et al., 2006a, 2006b). Then the effects of interventions are summarized as "DALYs averted" rather than "QALYs gained".[9]

For example, Canning (2006) assembles several evaluations of interventions that, in broad terms, either seek to prevent the transmission of HIV/AIDS in Africa or to treat it. Preventative interventions cost between $1-$21 per DALY averted, including mass media campaigns; and peer education, condom distribution, and treatment of sexually-transmitted diseases among commercial sex workers. Blood transfusion safety and drug-based prevention of mother-to-child transmission cost less than $50 per DALY averted, followed by voluntary counseling and testing, expanded condom distribution, and other interventions. Treatment with anti-retroviral drugs was considerably more costly per DALY averted. Even so, Canning (2006) notes that cost-utility ratios are naturally sensitive to the high (but declining) costs of some anti-retroviral drugs, once again highlighting the importance of conducting sensitivity analysis and considering the generalizability of older effectiveness and costs results to new settings.

2.5 Perspective of the Cost Analysis

Whose costs and outcomes—whether benefits, effects, or utility—should be reflected in the estimation of an NPV or CER? The objective of CBA and CEA is usually to inform choices about the allocation of society's scarce resources for the betterment of society's outcomes. This social perspective implies that costs and, if possible, effects should be measured from multiple standpoints: (1) governments, including agencies directly and indirectly involved in implementing the intervention; (2) non-governmental organizations or private firms; (3) social sector clients, such as students, patients, and their families; and (4) non-clients who nonetheless

[9] Among several methodological distinctions, DALYs differ in the exact quality weights applied, since QALY weights are survey-based attempts to determine "utilities" and DALY weights are determined by an expert panel. Earlier applications of DALYs often incorporated age weights in addition to the usual time discounting (placing highest weight on young adults) (Jamison et al., 1993). Subsequent analyses abandoned the arbitrary choices inherent to age-weighting (Musgrove and Fox-Rushy, 2006). See Drummond et al. (2005, p. 187) and Musgrove and Fox-Rushby (2006) for a further discussion of DALYs and comparison to QALYs.

receive benefits or effects. There is widespread agreement about the conceptual importance of adopting a social perspective in CBA (Boardman et al., 2011) and CEA (Levin and McEwan, 2001). Indeed, CEA guidelines in health policy adopt the social perspective as part of a standard "reference case" analysis (Gold et al., 1996).

Even so, cost analyses in CBA and CEA do not always adopt a social perspective. First, while it is common to include costs that accrue to an implementing agency, such as a Ministry of Education or Health, it is less common to include costs borne by families such as the opportunity cost of time devoted to travel or volunteer labor. This is often due to the practical difficulties of obtaining data on the value of "free" resources used in an intervention (Musgrove and Fox-Rushby, 1996), or because cost estimates rely exclusively on budget data that only report government agency costs.

Second, many social interventions involve transfer payments (e.g., school scholarships or conditional cash transfers). The transfer is a cost to the government or implementing agency, but a benefit to recipients, and it should be excluded from a social cost estimate in CEA (in a CBA, the cost and benefit simply cancel each other out).[10] If transfer payments are included, the analysis explicitly adopts a government perspective. Several Kenyan interventions cited in Table 1 involve a substantial component of transfer payments, and the final column illustrates how their inclusion in cost estimates can influence cost-effectiveness rankings.

The challenge of adopting a social perspective is even more apparent when estimating benefits or effects. Education and health interventions may create positive externalities: that is, benefits that accrue to non-clients such as untreated classmates of treated children. Externalities could also be negative and thus fall under costs. In an evaluation of school-based deworming, Miguel and Kremer (2004) found that interventions improved health and school participation among treated children, but also among untreated children in intervention schools and in neighboring non-intervention schools. In the U.S., the Perry Preschool Project created positive externalities for community members in the form of reduced criminal activity among treated individuals (Duncan et al., 2010). When the averted costs are monetized, they accounted for a substantial proportion of social benefits. Despite these examples, the evidence base on external effects is comparatively sparse and many authors focus exclusively on private benefits and

[10] The distribution of transfer payments does incur social costs. For example, Caldés et al. (2006) calculate that costs of subsidy distribution in Mexico's Progresa program were almost 5% of the total transfer amounts.

effects received by clients.

2.6 Ex Post versus Ex Ante CEA

In the academic literature, CEA is almost exclusively ex post, with the objective of identifying which of at least two interventions, X or Y, improved a specific outcome at least cost. In applied decision settings, such as a government or international organization, the CEA is often ex ante.[11] It is used to judge whether a hypothetical intervention, Z, *should* receive investments instead of other candidates such as X or Y. While a variant of Z might have been implemented and evaluated, it is possible that it only exists on the planner's drawing board.

For the moment, the distinction between ex post and ex ante CEA is less germane than it might seem. Both approaches are fundamentally comparative, and so it is important to gather and review the extant literature on the costs and effects of *any* candidate intervention—X, Y, or close relatives of Z—that pursued similar objectives. Naturally, some of this literature is bound to be of lower quality, and the next two sections highlight the most important considerations when gathering evidence on effects and costs, respectively. Section 7 will revisit the issue of ex ante CEA, focusing on the common case when data on the costs and/or effects of a candidate intervention Z are sparse.

[11] Ex ante CEA is rare in World Bank projects (World Bank, 2010), although the term is used frequently. In a similar vein, Clune (2002) showed that the term cost-effectiveness is predominantly rhetorical in U.S. education literature, and is rarely accompanied by a concrete empirical analysis.

3. Measuring the Effects of Interventions

3.1 What Makes a Good Impact Evaluation?

Social scientists apply two general criteria in judging the quality and usefulness of impact evaluations: internal and external validity (McEwan, 2008; Shadish, Cook, and Campbell, 2002). An estimate of effectiveness is internally valid when it identifies a credible causal link between an intervention and an outcome measure, in a particular sample of subjects.

The causal effect of an intervention is the difference between subjects' outcomes when treated by an intervention, and the same subjects' outcomes when not treated. The latter is called the counterfactual. (The subjects in question may be students, patients, or another unit of observation.) Short of procuring a time machine, the counterfactual cannot be observed because treatments cannot be undone. Instead, research methods are employed to "create reasonable approximations to the physically impossible counterfactual" (Shadish et al., 2002, p. 5).[12] Researchers estimate counterfactual outcomes by identifying a separate group of untreated subjects, called a control group. It provides a valid counterfactual to the extent that control subjects are similar to treated ones, on average, but for their exposure to the treatment. This is most likely to occur in a randomized experiment or a high-quality quasi-experiment.

An estimate of effectiveness is externally valid when it can be generalized to modified versions of the intervention, to different samples of subjects, and to different policy contexts. For example, conditional cash transfers programs have been implemented in many countries (Fiszbein and Schady, 2009). It is not always certain whether estimated effects in one program can be generalized to cases where the cash transfer is smaller, the target subjects are much younger or older, or contexts where a lower percentage of children already attend school prior to the intervention. These questions can often be directly addressed by conducting new impact evaluations, as the burgeoning literature on CCTs has shown. More often, especially for little-researched interventions, judgments about external validity are not clear-cut and are informed by common sense and theory. The particular issue of external validity in CEA is revisited in section 6.

[12] Shadish et al. (2002) describe the history of the counterfactual reasoning. Statisticians formalized the framework, especially Donald Rubin (Holland, 1986), in a framework that has been adopted within impact evaluation (e.g., Ravallion, 2005).

3.2 Methods of Estimating Effectiveness

The literature in impact evaluation focuses overwhelmingly on the importance of improving internal validity.[13] One of the greatest threats to internal validity is selection bias, or a pre-existing difference (such as poverty) between subjects in treatment and control groups. If the pre-existing differences between two groups cause differences in outcomes, then the control group provides a poor estimate of the counterfactual and any differences can be mistaken for an intervention's *effectiveness*. Researchers employ two approaches, often in combination, to ensure internal validity: (1) evaluation design and (2) statistical controls.

Evaluation design. Sound evaluation design, especially randomized experiments and good quasi-experiments, are the best means of ruling out selection as a threat to internal validity. In the classic randomized experiment, researchers flip a coin to determine which subjects are treated, and which are not (noting that student, patients, schools, clinics, or entire territorial units could also be randomly assigned). Thus, each subject's chances of receiving the treatment are identical.[14]

For example, the Honduran PRAF program awarded cash transfers to eligible families living in entire municipalities that were randomly assigned to participate in the program. The treatment group of families living in randomly treated municipalities was similar to a control group of families living in untreated municipalities, on average, but for its exposure to the treatment (Galiani and McEwan, 2011). Hence, any subsequent differences and education and health outcomes could be causally attributed to the treatment. The same logic underlies conditional cash transfer experiments in Mexico, Nicaragua, and elsewhere (Fiszbein and Schady, 2010).

In quasi-experiments, the broadest category of research, assignment may contain elements of randomness or purposeful assignment by the researcher, but some might be due to the individual choices of students, parents, or administrators. When greater control is exerted, then causal results often possess greater internal validity. One of the most credible quasi-

[13] This paper does not describe methods in detail. For extensive reviews, see Murnane and Willett (2011), McEwan (2008), Ravallion (2001, 2005), Angrist and Krueger (1999), Imbens and Wooldridge (2009), DiNardo and Lee (2010), and Shadish, Cook, and Campbell (2002).

[14] A coin flip or similar mechanism is only the simplest approach to designing randomized assignment. The essential point is that students or other units have well-defined probabilities of being assigned to the treatment. On the design, implementation, and analysis of randomized experiments, see Orr (1999) and Duflo, Glennerster, and Kremer (2006).

experimental methods is the regression-discontinuity design (RDD).[15] In the RDD, researchers assign subjects to treatment or control groups on the basis of a single assignment variable—often household income or a qualifying pre-test, but potentially any continuous variable—and a specified cutoff value. For example, Chile assigned entire schools to receive an intensive tutoring intervention if the mean test scores of the school fell below a specified cutoff value (Chay, McEwan, and Urquiola, 2005). Assignment was not randomized, as in the flip of coin, but neither was it primarily due to unobserved decisions of administrators. This provides sufficient leverage to identify the causal effect of tutoring on the later test scores of schools.

The causal effect is estimated by comparing the outcomes of treatment and control schools whose values of the assignment variables are just below or just above the assignment cutoff. The intuition is that schools should be very similar, not just in their values of the pre-test used to assign the program, but in other observed and unobserved ways.[16] In the tutoring example, control schools (just to the right of the cutoff) provide a good counterfactual estimate of outcomes for treated schools (just to the left). Thus, any sharp—or discontinuous—changes in school outcomes near the cutoff can be attributed to the tutoring treatment.[17]

Statistical controls. When evaluation design is not an option, researchers use statistical methods to control for observed pre-existing differences between treatment and control groups. (Even in well-designed experiments and quasi-experiments, controlling for pre-existing differences can be a useful means of reducing the standard errors of impact estimates.) Linear regression is the most popular method of statistically controlling for observed differences between subjects in treatment and control groups.

It is increasingly accompanied by methods relying on the propensity score (Imbens and Wooldridge, 2009). The propensity score is the predicted probability that a subject is treated, obtained from a logistic regression of a treatment dummy variable on independent variables

[15] See especially Lee and Lemieux (2010) and DiNardo and Lee (2010). Other quasi-experimental methods (e.g., "natural" experiments occasioned by variation in an instrumental variable) can vary substantially in their internal validity, often depending on the context.

[16] Lee and Lemieux (2009) interpret the RDD as a "local" randomized experiment, since the precise location of subjects on either side of the cutoff is partly due to random noise in the assignment variable. Of course, this interpretation falls apart if subjects can precisely manipulate their values of the assignment variable and, thus, their treatment status.

[17] Buddelmeyer and Skoufias (2003) compare experimental and RDD estimates of a single intervention: Mexico's Progresa program. While the original Progresa design included randomized assignment to localities, it also assigned households within localities on the basis of a poverty index and associated eligibility cutoff. They find close correspondence, suggesting that the internal validity of RDD results is high.

correlated with treatment group status (such as gender or poverty). It is a convenient index of observed differences between subjects. In the spirit of a quasi-experiment, researchers use propensity scores to match "similar" observations in treatment and control groups, or to re-weight treatment and control observations to attain balance in observed variables across treatment and control groups, thus mimicking a randomized experiment.[18] Unfortunately, both regression and propensity score methods hang their hats on an assumption: that selection to treatment and control groups is entirely determined by observed variables. In this sense, the methods are no panacea for selection bias introduced by unobserved differences between treatment and control groups, which often occurs in settings without a prospective evaluation design.

3.3. Comparing Effectiveness of Different Interventions

In the best circumstances, an impact evaluation will simultaneously estimate the effectiveness of several interventions, relative to a no-intervention control group. The randomized experiment of Muralidharan and Sundararaman (2011) evaluated the impact of several Indian interventions on students' test scores, including group teacher incentives, individual teacher incentives, block grants to schools, and additional teachers' aides. Given the common outcome measures applied to subjects in all treatment and control groups, it was straightforward to estimate and compare the magnitudes of each intervention's effect. The authors further calculated incremental costs of each intervention (relative to the control group) and assessed relative cost-effectiveness.

It is far more common—as in Table 1—that a CEA gathers estimates of effectiveness from several studies that use roughly similar outcomes (e.g., "mathematics achievement") but different tests and measurement scales. Thus, a point gained on one test cannot be meaningfully compared to a point gained on another. In such cases, the common practice is to report impacts as effect sizes, by dividing an impact estimate by the full-sample standard deviation of the respective outcome variable. The subfield of meta-analysis has developed a toolkit for the calculation of comparable effect sizes, in these and more difficult circumstances (Rosenthal,

[18] Imbens and Wooldridge (2009) show how the propensity score can be easily used to construct inverse probability weights. The weights, in turn, can be used to estimate a weighted mean difference in the outcomes of treatment and control subjects that controls for observed differences. Imbens and Wooldridge further demonstrate how regression can be combined with inverse-probability weighting, in weighted least-squares regressions. Then, the impact estimates are consistently estimated if *either* statistical model (the linear regression, or the logistic regression) is correctly specified. In this sense the estimates are "doubly robust."

1994). Evans and Ghosh (2008) present an example of a CEA that uses effect sizes in the calculations of multiple CERs. Even in such circumstances, one must be circumspect about interpreting results; especially tests are applied in different grades or reflect the content of heterogeneous curricula.

Another common scenario is that studies report effectiveness for an intermediate outcome measure or a final outcome measure, but not both. In such cases, authors commonly convert intermediate effects into final effects by assuming an unknown parameter or estimating it with further statistical modeling (often using a secondary data set). For example, JPAL (2011) compares the cost-effectiveness of multiple interventions in reducing the incidence of child diarrhea, a final outcome, although two experiments only report effects on an intermediate outcome: change in water chlorination rates. The CEA used descriptive data to inform its assumptions about the relationship between chlorination rates and eventual incidence of diarrhea.

In the health and medical literature, Drummond et al. (2005) and Gold et al. (1996) describe many examples in which secondary models are used to convert intermediate outcomes into final outcomes. For example, a short-term impact evaluation may reveal that drug therapies reduce cholesterol levels. However, regression modeling with separate data may be needed to estimate the (non-experimental) effect of cholesterol reduction on the incidence of coronary heart disease and mortality. The same authors describe the tools used to convert disparate effectiveness measures into estimates of QALYs gained (also see the textbook of Muennig, 2008). Musgrove and Fox-Rushby (2006) and Tan-Torres Edejer et al. (2003) describe related methods and guidelines for the estimation of DALYs.

In education CEA and CBA, it is increasingly common to use parameter estimates from secondary data analysis in order to convert effects on years of schooling attained or test scores into effects on lifetime earnings. In a CBA, Schultz (2004) converts the impact of Mexico's Progresa program on schooling attainment into effects on lifetime earnings. Using household survey data, he estimates the non-experimental effect of attainment on earnings, and then applies this result to experimental attainment effects obtained from the original evaluation. He reports an internal rate of return of 8%, where the only benefits are private earnings gains.

In the U.S., Levine and Zimmerman (2010) gather effectiveness data on a wide range of education and social interventions, often expressed as effects on test scores. Using secondary data and a literature review, they conclude that a 1 standard deviation increase in test scores

yields a 10 percent increase in earnings; this is a vital ingredient in CBAs of each intervention. This type of extrapolation is rare in developing countries, if only because few datasets contain both childhood test scores and adult earnings, which would allow researchers to estimate the (non-experimental) relationship between the two variables.[19]

The prior literature reveals a trade-off between internal validity and the potential richness of outcome measures in a CEA. Randomized experiments frequently report short-run effects over one or two years. Researchers can apply secondary data analysis of non-experimental data to convert intermediate outcomes to longer-run, final outcomes (or even facilitate a full CBA). Because these secondary exercises rely on non-experimental correlations, they threaten the internal validity of effects on long-run outcomes or benefits. The partial correlation between test scores and later earnings, for example, may not be causal. The best compromise is to report comparable CERs for interventions using the short-run measures of effectiveness, but also extrapolated CERs using longer-run estimates of effectiveness obtained with secondary modeling. The latter exercise should also be accompanied by additional sensitivity analysis, especially on the assumed or estimated parameters.

[19] Muralidharan and Sundararaman (2011) report an exception from India.

4. Measuring the Costs of Interventions

4.1. The Ingredients Method

This section describes an intuitive approach for estimating incremental costs called the ingredients method.[20] It relies on the identification of all resources or ingredients consumed in an intervention and the valuation of each ingredient. This, in turn, is used to estimate the incremental costs of the intervention. It is important to emphasize that the objective is to estimate incremental costs incurred over the duration of the intervention that was actually evaluated. In recent randomized experiments, this is often just a year, but it could be longer depending on the particular intervention.

Interventions use resources that have valuable alternative uses. For example, a program for raising student achievement requires personnel, facilities, and instructional materials. By devoting these resources to a particular activity we are sacrificing the gains that could be obtained from using them for some other purpose. Technically, then, the cost of a specific intervention will be defined as the value of all of the resources that it utilizes had they been assigned to their most valuable alternative uses. In most cases, the market price of a resource suffices, but for other resources, especially non-marketed ones like contributed facilities or labor, a shadow price must be estimated.

4.2. Specification of Ingredients

The specification of ingredients is often facilitated by dividing ingredients into categories including (1) personnel, (2) facilities, (3) equipment and materials, (4) other program inputs, and (5) client inputs. Other methods, such as activity-based costing, delineate a series of "activities" that comprise an intervention, and then categories of ingredients within each activity, but the spirit of the costing exercise is similar (e.g., Fiedler, Villalobos, and De Mattos, 2008).

Personnel. Personnel ingredients include the additional human resources required to implement the intervention. This category includes full-time personnel as well as part time employees, consultants, and volunteers. All personnel should be listed according to their qualifications and time commitments. Qualifications refer to the nature of training, experience, and specialized skills required for the positions. Time commitments refer to the amount of time

[20] For similar descriptions in education and health, respectively, see Levin and McEwan (2001, Chapters 3-5) and Drummond et al. (2005, Chapter 4).

that each person devotes to the intervention in terms of percentage of a full-time position. In the latter case there may be certain employees, consultants, and volunteers who allocate only a portion of a work-week or year to the intervention.

Fiedler et al. (2008) conducted a cost analysis (but not a CEA or CBA) of a Honduran intervention that used volunteer community monitors to assist mothers in monitoring the growth and health of young children. The authors specified in detail the activities of the monitors and the number of hours per month associated with activity, including monthly weighing sessions attended by mothers and monthly follow-up visits to households. They further specified the qualifications and time commitments of personnel involved in training and supervising the monitors, including physicians, nurses, nurses' aides, health educators, drivers, and so on. The key point that is the study exhaustively described the incremental personnel required to faithfully implement the intervention, whether volunteers or not.[21]

Facilities. Facilities include any classroom or clinic space, offices, and other spaces used by the intervention, including donated ones. All such requirements must be listed according to their dimensions and characteristics, along with other information that is important for identifying their value, such as quality of construction. Any facilities that are jointly used with non-intervention activities should be identified according to the portion of use that is allocated to the intervention.

Equipment and materials. These refer to furnishings, instructional equipment, and materials that are used for the intervention, whether covered by project expenditures or donated. Specifically, they include classroom and office furniture, instructional equipment as computers, books and other printed materials, office machines, paper, and other supplies. Both the specific equipment and materials solely allocated to the intervention and those that are shared with other activities should be noted.

Other inputs. This category refers to all other ingredients that do not fit readily into the categories set out above. For example, it might include insurance, telephone service, electricity, heating, internet access, and so forth. Any ingredients that are included in this category should be specified clearly with a statement of their purpose.

Required client inputs. This category of ingredients includes any contributions that are

[21] A common shortcut strategy is to simply use a budget line item for personnel. However, this risks overlooking personnel paid out of a separate budget (but who nonetheless contribute time to the intervention). It also misses volunteer personnel who receive no wages.

required of the intervention's clients or their families. For example, if an education intervention requires the family to provide transportation, books, uniforms, equipment, food, or other student services, these should be included under this classification. To provide an accurate picture of the social costs of replicating an intervention that requires client inputs, it is important to include them in the analysis.

In the Honduran example, Fiedler et al. (2008) acknowledge that households incur time-related costs by participating in the program. However, they make the explicit decision to exclude client ingredients and limit the perspective of the cost analysis to the government. To be fair, this is by far the most common scenario in developing-country cost analyses, although it tends to underestimate full social costs. It may skew cost-effectiveness comparisons when an intervention relies heavily on stakeholder time to effective implementation. For example, El Salvador's EDUCO program decentralized responsibility for managing local schools to groups of parents and community members (Jimenez and Sawada, 1999).

Despite the importance of accounting for ingredients, it would be unwise to let perfect be the enemy of good, especially in cost analyses conducted with limited time or resources. The degree of specificity and accuracy in listing ingredients should depend upon their overall contribution to the total cost of the intervention. For example, personnel inputs represent three-quarters or more of the costs of education interventions, although situations vary by intervention. The important point is that an eventual error of ten percent in estimating personnel costs will have a relatively large impact on the total cost estimate because of the importance of personnel in the overall picture. However, a 100 percent error in office supplies will create an imperceptible distortion, because office supplies are typically an inconsequential contributor to overall costs.

4.3. Sources of Information

Information on the ingredients used in an intervention can be obtained in three ways: (1) through document review, (2) through interviews with personnel or observations of the intervention, and (3) through empirical analysis. An essential starting point is the examination of program documents. These documents include general descriptions of the program prepared by program staff or outsiders, budgets and expenditure statements, web sites, and reports by previous evaluators of the program.

A second source of information is interviews with individuals involved in the intervention. These individuals include designers; directors and administrative staff; front-line personnel such as teachers and doctors; and clients. Even after conducting interviews, it is often helpful to conduct direct observations of the intervention. In a school intervention, for example, the evaluator might sit in on several classes. The purpose of doing so is to ascertain if the prescribed ingredients are actually being used. If the program designer mentioned that students should have individual workbooks, is it the case that all students in the class have workbooks? If program documents state that 50 minutes of classroom time is devoted to instruction, is this revealed during classroom observations?

The concern is not merely academic, since incomplete and variable program implementation is a common occurrence, even in "best-practice" education programs (e.g., McEwan, 2000; Loeb and McEwan, 2006). In the U.S., Levin et al. (2007) document that costs of adolescent literacy programs can vary widely across sites due to variable program implementation, even though each site implements a nominally homogeneous "program." Their analysis illustrates the potential biases of cost estimates that rely exclusively on back-of-the-envelope calculations using program documents.

A third source of information is empirical analysis of intervention data, ideally collected in the impact evaluation itself. Randomized experiments often collect extensive background data from clients (as household surveys or individual questionnaires) and providers (questionnaires filled out by school principals or physicians). The surveys often contain rich data on the qualifications, time use, and remuneration of stakeholders involved in the intervention. The virtues of using empirical data are twofold: (1) they are often collected for both treatment and control groups, allowing for a more careful identification of incremental ingredients consumed, and (2) they are less likely to overstate ingredient usage, since they reflect the intervention as it was implemented, rather than designed.

4.4 Valuing Ingredients

At this second stage a value is placed on each ingredient, regardless of whether or not it is found in a budget document. Personnel costs are relatively easy to estimate by combining salaries and benefits (although the latter are frequently ignored or underestimated, particularly for civil servants, because benefits do not appear in program budgets). Ultimately, they should include the

full value of what it takes to obtain persons with the desired qualifications. In the Honduran example, Fiedler et al. (2008) faced the dilemma that community health monitors were actually volunteers. They received minimal in-kind remuneration, which the authors costed out at less than 10 cents per working hour, although they received no additional wages. To estimate the full social costs of the intervention, one option would be to estimate the value of the contributed services by using secondary data such as a household survey, and the market wages of female workers with similar levels of formal schooling and experience.

Incremental facilities costs are usually more of a challenge because many education and health organizations already own their facilities, so it is not always obvious how to value facilities used in, say, a one-year intervention. One approach is to estimate how much it would cost to lease the facility for the duration of the intervention. A second is to amortize the total value of the facility over its remaining useful life, estimating an annual cost of the facility. The annualized value of a facility comprises the cost of depreciation (that is, how much is "used up" in a given year of a facility with a fixed life) and the interest forgone on the undepreciated portion.

A common assumption is that a facility with a thirty-year life, for example, loses one-thirtieth of its value each year in depreciation cost. Furthermore, since the undepreciated investment cannot be invested elsewhere, it implies an additional cost of forgone interest income. When equipment and materials (e.g., lab equipment or textbooks) have a usable life of more than one year, those costs can also be annualized. For example, a cost-effectiveness analysis of interventions to improve test scores among Brazilian students interviewed teachers and school personnel about the useful life of infrastructure investments, and used these data to annualize costs (Harbison and Hanushek, 1992). The concepts and formulas for annualization are discussed in Levin and McEwan (2001) and Drummond et al. (2005).

4.5. Adjusting Costs for Inflation, Time-Value, and Currency

Inflation. In many interventions, the effects and costs are measured over a time period of one year. In others, the costs are incurred over several years. In the latter case, the cost analysis must adjust costs for price inflation, assuming that the prices of ingredients are expressed in nominal terms (that is, prices specific to each year of an intervention). Suppose that the total ingredients cost of an intervention is $100 in 2005, the first year, and $120 in 2006, with each cost expressed

in nominal prices. Price inflation is summarized in a price index such as the consumer price index or the GDP deflator. Suppose the index value is 120 in 2005 and 125 in 2006, indicating price inflation of 4.2% [(125-120)/120]. The goal is to express each cost in real 2005 dollars. The 2005 value obviously does not change, and the 2006 cost in real 2005 dollars is $115.20 [(120/125)x$120].

Discounting. Even if inflation is zero, society is clearly not indifferent between a $100 cost incurred now and a $100 cost incurred several years in the future. To make them comparable, future costs are discounted by the formula $\sum_{t=0}^{n} \frac{C_t}{(1+r)^t}$, where C is the incremental cost incurred in year t (with t=0 indicating an immediate cost and n being the final year of the intervention), and r is the discount rate. When the CBA or CEA is conducted from society's perspective, it is commonly interpreted as a social discount rate (SDR), and an extensive theoretical and empirical literature informs its choice (see, e.g., Boardman et al., 2011).

From an applied standpoint, two issues are important. First, the SDR is often mandated by the organization sponsoring the CEA. The Inter-American Development Bank applies a consistent rate of 12%; Dhaliwal et al. (2011) summarize the discount rates used by other governments and agencies. Second, a CEA is fundamentally comparative and will often compare the CERs from several studies (as in Table 1). To ensure comparability, it is important that CERs use a similar set of costing assumptions, including the discount rate.

Currency. Once adjusted for inflation and time, the analyst may wish to report costs in local currency if the CEA is conducted on two or more interventions implemented in that country. If the CEA compares CERs for interventions conducted in different countries, authors sometimes use nominal exchange rates to convert costs to a common currency (typically U.S. dollars). However, poorer countries tend to have relatively lower prices for nontraded goods and services, and local currency actually has more purchasing power than it would in global markets. Purchasing power parity exchange (PPP) rates, based on extensive international price surveys, adjust for these differences (World Bank, 2008). Given the substantial variance in how authors deal with currency conversions, it is sensible to report dollar costs obtained using both nominal and PPP rates.

5. Who Does CEA in Developing Countries?

5.1 Overview

A review of World Bank projects gloomily concluded that CEA is often invoked as a method, but is rarely applied in a rigorous attempt to compare the incremental effects and costs of two or more interventions. Of 24 projects that purported to conduct a CEA, only one actually did (World Bank, 2010).[22] It is reasonable to conclude that ex ante CEA in education and health projects is rare and, when applied, is often misconstrued as a simple cost analysis (excluding consideration of effects), or as a CBA-type method capable of judging the potential worth of a *single* intervention.

The ex post CEA literature justifies more optimistic conclusions. While not common, the use of CEA in education and health evaluations has grown in the last decade. The emerging literature falls into three broad categories. In the first, CEA is embedded within a non-experimental impact evaluation. More recently, randomized experiments have embedded CEA. Finally, there are ambitious attempts to gather CERs from a wide array of studies, and to construct "league tables" that can provide general guidance on resource allocation.

5.2 CEA Embedded in Non-Experimental Impact Evaluations

Before the randomization boomlet of the last decade, education economists in developing countries were mainly engaged in two activities: (1) estimating the rate of return (ROR) to increased years of schooling using non-experimental Mincer earnings regressions, and (2) estimating the marginal effect of increasing various school inputs (e.g., textbooks, teacher salaries) on student achievement and attainment), using non-experimental education production functions (EPF).

ROR Studies. The first approach yielded hundreds of estimates of the rates of return to various levels of schooling (Patrinos and Psacharopoulos, 2010; Jimenez and Patrinos, 2008). These studies constituted an important application of CBA, and influenced the terms of debate about the importance of human capital investment. They also provide a rich source of secondary estimates of the schooling-earnings relationship, which allows effects on attainment to be

[22] In the U.S. education literature, Levin and McEwan (2001, 2002) were similarly pessimistic about quantity and quality of CEA. The health CEA literature in the U.S. is more robust, with several textbooks (Drummond et al., 2005; Muennig, 2008; Gold et al., 1996) and numerous cost-effectiveness comparisons of health and medical interventions.

transformed into benefits (Schultz, 2004; Levine and Zimmerman, 2010).

However, rate-of-return studies were challenging to use for resource allocation. First, they were not evaluations of well-specified education interventions, but rather the average impact of increasing the quantity of any formal schooling. Second, the non-experimental estimates may have suffered from selection bias, given the correlation of schooling with a variety of unobserved determinants of earnings, such as ability.[23] Duflo (2001) notably addressed both challenges in a CBA, by evaluating a specific intervention—Indonesia's ambitious school construction campaign in the late 1970s—which provided a credibly exogenous source of variation in school supply. She found that each new school constructed per 1000 children increased schooling by 0.12-0.19 years, implying economic returns of 6.8-10.6 percent.

EPF Studies. Second, education economists estimated non-experimental education production functions, regressing student achievement or another outcome on a "kitchen sink" of student, family, teacher, and school variables (for reviews, see Glewwe and Lambert, 2010). The regression studies yielded many estimates of the marginal impact of resources on student outcomes, which can be conveniently interpreted as the denominator of the CER in section 2. However, they had two shortcomings. First, most studies reported no estimates of incremental costs of the inputs, so CERs could not be calculated. Second, the causal interpretation of the regression coefficients on school resources was not always clear, particularly given the lack of ex ante evaluation designs.

However, a few notable EPF studies combined cost data with coefficient estimates to conduct a full CEA.[24] For example, Harbison and Hanushek (1992) estimated the effects of school inputs on language scores of Brazilian second-graders, holding constant family background. They also calculated incremental input costs per student, annualizing capital costs, and calculated CERs. Among other findings, textbooks cost only $0.26 to raise scores by one point, compared with the $6.50 in teacher salaries required to obtain the same increase in test scores.

[23] Literature in developed countries focused extensively on methods to correct for selection bias, including twins studies and instrumental variables estimates (see Gunderson and Oreopoulos, 2010 for a review). The developing-country literature is surprisingly weak in this area, although Patrinos and Psachropoulos (2010) review some exceptions.

[24] Other notable non-experimental studies include World Bank (1997) on India; Bedi and Marshall (1999) on Honduras; Fuller et al. (1994) on Botswana; Glewwe (1999) on Ghana; and Tan et al. (1997) on the Philippines.

5.3 CEA Embedded in Experimental Impact Evaluations

Randomized experiments have improved greatly on the internal validity of non-experimental regression studies and often apply CEA, ranging from back-of-the-envelope cost calculations to thorough attempts to assess cost-effectiveness. Many of latter studies were conducted in India and Kenya in the last decade, including the Kenyan studies summarized in Table 1.[25] In an early example, Miguel and Kremer (2004) find that a Kenyan school-based deworming intervention increases school participation by 0.14 year per treated child, and costs about $3.50 per year gained. The cost per year of related interventions, like providing free school uniforms, was substantially higher. Because the intervention also affected health outcomes, the authors calculated that it reduced DALYs by 649, or $5 per DALY averted, which compares favorably with other options. Finally, they conducted a partial CBA by using a ROR study to estimate the effects of increased school participation on lifetime earnings. They find that deworming increase the present value of wages by $30 per treated child, at a cost per child of $0.49.

The study reiterates three issues in the conduct of CEA. First, human capital interventions often affect multiple outcomes measures in education and health, and the CEA should cast as a wide a net as possible in measuring outcomes and calculating CERs. If possible, the analysis should conduct a limited CBA by attempting to convert one of more effectiveness measures into lifetime earnings benefits, using ROR estimates obtained from the literature or supplemental analysis of household survey data.

Second, even randomized evaluations are forced to rely on assumptions and non-experimental parameter estimates to calculate useful measures of effectiveness or benefits (illustrated by the DALY calculations and CBA). As in Miguel and Kremer (2004), the assumptions and methods need to be carefully stated and sensitivity analysis should be conducted.

Third, contemporary impact evaluations devote the lion's share of attention to effects, with relatively less attention to costs. Even the deworming study uses a cost estimate from a Tanzanian program reported in the secondary literature. It would be ideal to report cost estimates of the intervention evaluated, along with some breakdown of ingredients, costs, and related assumptions.

[25] Other studies include Muralidharan and Sundararaman (2011) on teacher merit pay; Kremer et al. (2004, 2009) on incentive payments to students; and Banerjee et al. (2005, 2007) on computer-assisted learning and tutoring.

5.4 CEA League Tables

Education Interventions. The growth in CEA has facilitated attempts to combine CERs from many studies in so-called league tables. Lockheed and Hanushek (1988) provide a pioneering example for education interventions, but they necessarily emphasized non-experimental evaluations. More recently, Evans and Ghosh (2008) utilize a broad base of experimental and non-experimental studies that attempt to measure impacts on school enrollment, school participation, and test scores. The authors standardize effectiveness measures when appropriate (e.g., converting test score effects to effect sizes). Costs are converted to annualized costs when appropriate, or a present value when an intervention is evaluated over several years. Finally, they convert costs to 1997 U.S. dollars and use nominal exchange rates (they considered using purchasing power parity exchange rates, but were hampered by the fact that many authors simply report costs in U.S. dollars with few details about local currency costs or exchange rate conversions). Finally, MIT's Poverty Action Lab reports CERs for interventions evaluated with experimental methods that share an outcome measure, notably student attendance and teacher attendance.[26] The assumptions used in their CEA are described by Dhaliwal et al. (2011).

Health interventions. International league tables for health interventions commonly adopt the DALY as a measure of effectiveness. The most ambitious effort to date in the second *Disease Control Priorities in Developing Countries* (Jamison et al., 2006a, 2006b), which commissioned reviews of interventions' effectiveness and costs and summarized the cost per DALY averted (Laxminarayan et al., 2006). The validity of these comparisons depends on the use of common methods to calculate costs and DALYs, described in Musgrove and Fox-Rushby (2006).

Comparisons across studies are considerably more difficult when they make different assumptions about discount rates applied to costs and DALYs, appropriate categories of cost ingredients, and so on. Canning (2006) summarized multiple estimates of the cost-effectiveness of HIV/AIDs prevention and treatment, finding that the cost per DALY averted is lower for prevention-related interventions than treatment-related ones. CERs are comparable *within* each study he reviews, and the relative cost-effectiveness ranking of interventions is roughly consistent within each study. However, the CERs could be compared across studies because of varying methodological assumptions about discount rates, price levels, etc. Finally, the Poverty

[26] See http://www.povertyactionlab.org/policy-lessons, accessed on November 30, 2011.

Action Lab reports basic league tables for the cost-effectiveness of increasing healthcare provider attendance and reducing child diarrhea, again by focusing on only the highest-quality impact evaluations, and making a strong attempt to compare costs using a common set of assumptions and methods.

6. Issues in Conducting CEA

6.1 Sensitivity Analysis

The proliferation of league tables and attempts to compare CERs within and across studies should be accompanied by sensitivity analysis.[27] The most common approach is a simple one-way sensitivity analysis, in which an uncertain parameter is varied and the CER is re-calculated. The concern is that plausible changes in parameter values will substantially alter the cost-effectiveness ranking of interventions.

First, the discount rate is an obvious candidate for sensitivity analysis, since the assumed rate can vary substantially among researchers, governments, and international organizations. Whatever the standard applied within an organization, it is almost certain that a CEA will compare its results to CERs from other interventions that apply a particular discount rate.

Second, ingredients-based cost analysis can be subject to considerable uncertainty, since estimates are sometimes cobbled together from interviews, secondary data analysis, or observation. In too many cases, the incremental cost is reported as a "back-of-the-envelope" estimate with no supporting empirical evidence. It makes sense in this case to conduct sensitivity analysis on the quantity of costly ingredients or the price of intensively-used ingredients. In practice, this often means the quantity and price of labor ingredients, and the quantity, price, and assumed lifetime of facilities ingredients.

Third, one-way sensitivity analysis can be applied to the estimates of effectiveness, perhaps using the 95% confidence interval of the effect size as a very cautious range over which to re-calculate CERs. Many impact evaluations report effects by subgroups such as income, race, baseline pre-test, or, as in Table 1, by geographic area. In such cases, it is useful to report separate CERs for subgroups. The latter can also be helpful for gauging the external validity of results.

Finally, when varying a large number of parameters, it may also be helpful to conduct a "best case, worst case" analysis, in which an upper-bound CER reflects pessimistic assumptions about incremental costs and effects, and a lower-bound reflects optimistic assumptions.

[27] See Boardman et al. (2011) and Levin and McEwan (2001) for further discussion.

6.2 External Validity of Cost-Effectiveness Ratios

In the simplest case, CERs are used to inform resource allocation decisions in the original setting of the impact evaluation. It is far more common that CEA informs decisions to implement modified versions of the treatment (perhaps scaled up), among heterogeneous subjects, perhaps located in a different region or country. The basic question is whether incremental costs and effects can be generalized outside the context of the initial evaluation.

Interventions. The highest-quality impact evaluations are often randomized evaluations of small-scale interventions. One concern is that implementing the interventions on a larger scale will modify the incremental effects or costs per subject. Economists draw a distinction between partial-equilibrium effects observed in small-scale evaluations and general-equilibrium effects observed in larger-scale implementations. For example, one could imagine that a conditional cash transfer plan with expanded eligibility and larger transfers could affect local prices by stimulating demand for some goods and services (though the limited evidence from Progresa suggests otherwise; see Fiszbein and Schady, 2009).

A classic example in the U.S. is the Tennessee STAR class size experiment conducted in 79 elementary schools in the mid-1980s. The evaluation literature found effect sizes of approximately 0.2 standard deviations on test scores in the early grades (Schanzenbach, 2007). In the mid-1990s, the evaluation informed a massive, statewide implementation of class size reduction in California, which immediately created 25,000 new teaching positions. In competitive labor markets, one might expect that the increased demand would drive up teacher salaries. However, the immediate effect in higher-paying school districts was that many new positions were filled with certified teachers poached from lower-paying school districts. Other schools experienced a sharp increase in the number of uncertified teachers (now staffing smaller classes). Jepsen and Rivkin (2009) suggest that the general equilibrium result, not apparent from the small-scale experiment, was to trade off positive effects of class size reduction against negative effects of inexperienced teachers in those classrooms.

Subjects. Impact evaluations typically report average treatment effects that can be generalized to the population from which the sample was drawn. Sometimes the population is well-defined and broad. For example, Muralidharan and Sundararaman (2011) drew a large, representative sample of schools in an Indian state, and only then proceeded to randomly assign schools to several interventions and a control. In contrast, many experiments begin with a

convenience sample of individuals, schools, or clinics that were solicited to participate in the experiment. In such cases, it is not clear under what conditions the effects and costs obtained in a "volunteer" sample can be generalized to a broader population.

There is no recipe for ensuring external validity (for a discussion, see Shadish, et al. 2002). One option is to enrich the CEA with estimates drawn from non-experimental impact evaluations that nonetheless reflect a representative sample of the relevant population. However, doing so often trades off internal validity for generalizability. A second is to examine the original impact evaluation for evidence of heterogeneous effects across types of subjects (often initial levels of human capital, as gauged by an intake assessment or level of poverty). The effects obtained from one subgroup may be considered more relevant, and emphasized in the calculation of CERs. For example, McEwan (2010) reports estimates of the effects of higher-calorie meals on student outcomes in Chile, but focuses on the subpopulation of poorer, rural students whose effects may be more generalizable to lower-income populations of Latin America.

Settings. Finally the setting of an intervention can have great bearing on whether effects and costs can be generalized. Baseline conditions often vary substantially from country to country. In Mexico's Progresa evaluation, the net enrollment rate is over 90 percent among young, primary-aged children, and relatively small enrollment effects were observed for that group. In a Honduran evaluation, where net enrollments are much lower, the effects were larger.[28]

Finally, the quality of existing service provision in a setting can also mediate the observed effectiveness of interventions. Vermeersch and Kremer (2005) find that schools meals affect students' academic achievement, but only when teachers are more experienced. In the case of conditional cash transfers, one could similarly imagine that effects on final health and education outcomes of participating children and adults would be enhanced if the services were higher-quality.

6.3 Steps in Conducting a CEA

The objective of a CEA, whether ex post or ex ante, is to establish credible CERs for competing interventions X and Y. Ex ante CEA has the further challenge of assessing the comparative cost-effectiveness of a hypothetical intervention Z. Both types of analyses should begin with a

[28] See Galiani and McEwan (2011) and Fiszbein and Schady (2009, Table 5.1).

comprehensive literature review, conducted in several steps. (In fact, these steps closely parallel the process of constructing a credible CEA league table.) First, researchers should identify competing interventions that include but are not necessarily limited to X or Y. Ideally these should include popular or widely implemented interventions in similar contexts, and interventions with good impact evaluations.

Second, researchers should locate estimates of effectiveness for interventions, evaluate their internal validity, and possibly discard lower-quality studies. There are no hard-and-fast quality standards to follow. MIT's Poverty Action Lab exclusively reviews randomized experiments in its CEAs (Dhaliwal et al., 2010). The U.S. "What Works Clearinghouse", which reviews U.S. studies of education interventions, has more ecumenical guidelines (WWC, 2008). While they also prize randomized experiments, WWC standards concede that some quasi-experiments can meet standards of high-quality evidence, especially regression-discontinuity designs. WWC standards also allow evidence from randomized experiments to be discounted if a study is severely compromised by a methodological flaw such as subject attrition from treatment and control groups. Depending on the intervention, it is possible that high-quality studies are scarce, and the CEA could incorporate effect estimates from non-experimental estimates, with appropriate caveats (such as the early example of Harbison and Hanushek, 1992).

Third, researchers should ensure that effect estimates from various studies are in comparable units. In some cases, this might involve transforming a test score impact into standard deviation units, described previously. In others, it could involve transforming an intermediate outcome measure into a final outcome measure using auxiliary assumptions or data (JPAL, 2011). Health economics has a far more comprehensive methodological toolkit to inform such judgments when they concern transforming measured health outcomes into QALYs gained or DALYs averted. In education economics, techniques are quite variable across studies. Whatever the method, it should be explained in sufficient detail so that readers can replicate it.

Fourth, researchers should locate incremental cost estimates of each intervention, insofar as they are available. In the best circumstances they are reported in the evaluation. More likely is that they are located in the unpublished grey literature of working papers or commissioned reports. It is possible that no cost estimate exists in which case researchers might attempt to either (1) reconstruct an ingredients-based estimate using documents, interviews, and secondary data analysis, or (2) use a cost estimate of a comparable intervention from another setting (as in

Kremer and Miguel, 2004). Both should be accompanied by appropriate caveats and, eventually, sensitivity analysis.

Fifth, researchers should ensure that incremental costs from different studies are in comparable units, appropriately adjusting them for inflation, time value, and currency. It is quite likely at this stage that compromises will be made. For example, many authors report costs in dollars, and provide few details on exchange rates used to convert from local currencies.

Sixth, the research should calculate CERs using the preferred estimates of incremental effects and costs, and also conduct sensitivity analysis of CERs using a range of plausible assumptions about effects and costs. The range of reported sensitivity analysis should be inversely related to the quality of underlying estimates of costs and effects. For example, the cost estimate may omit categories of ingredients, or have uncertain provenance, necessitating a more cautious approach to sensitivity analysis.

Lastly, the ex ante CEA faces the challenge of comparing (known) CERs of X and Y to the (unknown) CER of a hypothetical intervention Z. Almost every development project includes the raw materials to construct an ex ante incremental cost estimate for a proposed intervention, including many ingredients and their prices. The main challenges at this stage are twofold. First, one should resist the urge to ignore ingredients that do not appear in project budgets, such as volunteer labor, unless the analysis explicitly adopts a government or funder perspective. Second, one should take care to express the incremental cost estimate in units comparable to those of CERs for X and Y, adjusting for time value, price, and currency when appropriate.

As a final step, the analyst should judge the potential cost-effectiveness of Z, by conducting a simple bounding exercise. Suppose that the known CERs of interventions X and Y are, respectively, $500 per unit of effect and $200 per unit of effect. The ex ante cost analysis reveals that a new intervention Z has a projected incremental cost of $1000, implying that its effect would have to be at least 5 units in order for its CER ($1000/5=$200) to equal the lowest observed CER. This usefully re-frames the resource allocation question to the following: is it likely that intervention Z will yield an effect of *at least* 5 units? To judge whether it likely, analysts may further consult the opinions of experts, qualitative evaluation literature, and impact estimates of related interventions that bear some resemblance to Z.

7. Conclusions

CEA is a useful but under-used tool to inform resource allocation decisions for a range of education and health interventions that share common measures of effectiveness. Even if analysts make the decision to conduct a full CBA, it is still possible to a conduct a CEA if there are credible estimates of incremental costs and effects for at least two interventions. Fortunately, the costs of conducting comparative CEA are likely to fall as the quantity and quality of impact evaluations rises in developing countries. Not all of those evaluations report incremental costs, and some that do treat it as a back-of-the-envelope exercise. Perhaps the greatest challenge moving forward is to systematize methods for calculating and comparing incremental costs, in much the same way that impact evaluators have done for the measurement of effectiveness.

References

Angrist, J. D., and A. B. Krueger. 1999. "Empirical Strategies in Labor Economics." In O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics* (Vol. 3A). Amsterdam: Elsevier.

Banerjee, Abhihit V., S. Cole, E. Duflo, and L. Linden. 2005. "Remedying Education: Evidence from Two Randomized Experiments in India." NBER Working Paper 11904. National Bureau of Economic Research.

Banerjee, Abhihit V., S. Cole, E. Duflo, and L. Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics.*

Bedi, A. S., and J. H. Marshall. 1999. "School Attendance and Student Achievement: Evidence from Rural Honduras." *Economic Development and Cultural Change* 47(3): 657-682.

Boardman, Anthony E., D. H. Greenberg, A. R. Vining, and D. L. Weimer. 2011. *Cost-Benefit Analysis: Concepts and Practice* (4th ed.). Boston: Prentice Hall.

Bouillon, César Patricio, and L. Tejerina. 2006. "Do We Know What Works? A Systematic Review of Impact Evaluations of Social Programs in Latin America and the Caribbean." Sustainable Development Department, Poverty and Inequality Unit. Washington, DC: Inter-American Development Bank,

Buddelmeyer, H., and E. Skoufias. 2003. "An Evaluation of the Performance of Regression Discontinuity Design on Progresa." Discussion Paper 827. IZA.

Caldés, Natalia, D. Coady, and J. A. Maluccio. 2006. "The Cost of Poverty Alleviation Transfer Programs: A Comparative Analysis of Three Programs in Latin America." *World Development* 34(5): 818-837.

Canning, David. 2006. "The Economics of HIV/AIDS in Low-Income Countries: The Case for Prevention." *Journal of Economic Perspectives* 20(3): 121-142.

Chay, Kenneth Y., P. J. McEwan, and M. S. Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools." *American Economic Review.*

Clune, William H. 2002. "Methodological Strength and Policy Usefulness of Cost-Effectiveness Research." In Henry M. Levin and Patrick J. McEwan, eds., *Cost-Effectiveness and Educational Policy: 2002 Yearbook of the American Education Finance Association.* Larchmont, NY: Eye on Education.

Dhaliwal, Iqbal, E. Duflo, R. Glennerster, and C. Tulloch. 2011. "Comparative Cost-Effectiveness to Inform Policy in Developing Countries." Unpublished manuscript, Abdul Latif Jameel Poverty Action Lab, MIT.

DiNardo, John, and D. S. Lee. 2010. "Program Evaluation and Research Designs." In Orley Ashenfelter and David Card, (Eds.), *Handbook of Labor Economics* (Vol. 4). Amsterdam: Elsevier.

Drummond, Michael F., M. J. Sculpher, G. W. Torrance, B. J. O'Brien, and G. L. Stoddart. 2005. *Methods for the Economic Evaluation of Health Care Programmes* (3rd ed.). Oxford: Oxford University Press.

Duflo, Esther. 2004. "Scaling Up and Evaluation." Paper prepared for the Annual World Bank Conference on Development Economics.

Duflo, E., R. Glennerster, and M. Kremer. 2006. *Using Randomization in Development Economics Research: A Toolkit.* Unpublished manuscript, MIT.

Duflo, Esther, and M. Kremer. 2005. "Use of Randomization in the Evaluation of Development Effectiveness." In George Keith Pitman, Osvaldo N. Feinstein, and Gregory K. Ingram, (Eds.), *Evaluating Development Effectiveness.* New Brunswick, NJ: Transaction Publishers.

Duncan, Greg J., Jens Ludwig, and K. A. Magnuson. 2010. "Child Development." In Phillip B. Levine and David J. Zimmerman (Eds.), *Targeting Investments in Children: Fighting Poverty When Resources are Limited.* Chicago: University of Chicago Press.

Evans, David K., and A. Ghosh. 2008. "Prioritizing Educational Investments in Children in the Developing World." RAND Labor and Population Working Paper WR-587. Santa Monica, CA: RAND.

Fiedler, John L., C. A. Villalobos, and A. C. De Mattos. 2008. "An Activity-Based Cost Analysis of the Honduras Community-Based, Integrated Child Care (AIN-C) Programme." *Health Policy and Planning* 23: 408-427.

Fiszbein, Ariel, and N. Schady. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty.* Washington, DC: World Bank.

Fuller, B., H. Hua, and C. W. Snyder. 1994. "When Girls Learn More Than Boys: The Influence of Time in School and Pedagogy in Botswana." *Comparative Education Review* 38(3): 347-376.

Galiani, Sebastian, and P. J. McEwan. 2011. "The Heterogeneous Impact of Conditional Cash Transfers in Honduras." Unpublished manuscript, Washington University and Wellesley College.

Glewwe, Paul, M. Kremer, and S. Moulin. 1997. "Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya," unpublished working paper.

Glewwe, Paul. 1999. *The Economics of School Quality Investments in Developing Countries: An Empirical Study of Ghana.* London: St. Martin's.

Glewwe, Paul, N. Ilias, and M. Kremer. 2003. "Teacher Incentives". NBER Working Paper 9671. National Bureau of Economic Research.

Glewwe, Paul, M. Kremer, S. Moulin, and E. Zitzewitz. 2004. "Retrospective vs. Prospective Analysis of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74(1): 251-268.

Glewwe, Paul, and S. Lambert. 2010. "Education Production Functions: Evidence from Developing Countries." In Dominic Brewer and Patrick J. McEwan, (Eds.), *Economics of Education.* Amsterdam: Elsevier.

Gold, Marthe R., Joanna E. Siegel, Louise B. Russell, and Milton C. Weinstein (Eds.). 1996. *Cost-Effectiveness in Health and Medicine.* New York: Oxford University Press.

Gunderson M., and P. Oreopoulos. 2010. "Returns to Education in Developed Countries." In Dominic Brewer and Patrick J. McEwan, eds., *Economics of Education.* Amsterdam: Elsevier.

Harbison, R. W., and E. A. Hanushek. 1992. *Educational Performance of the Poor: Lessons from Rural Northeast Brazil.* Oxford: Oxford University Press.

Holla, Alaka, and M. Kremer. 2009. "Pricing and Access: Lessons from Randomized Evaluations in Education and Health." Working Paper 158. Washington, DC: Center for Global Development.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association 81*(396): 945-960.

Imbens, Guido W., and J. M. Wooldridge. 2009. "Recent Development in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5-86.

Jamison, Dean T., W. H. Mosley, A. R. Measham, and J. L. Bobadilla, eds. 1993. *Disease Control Priorities in Developing Countries.* New York: Oxford University Press.

Jamison, Dean T., Joel G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, Prabhat Jha, Anne Mills, and Philip Musgrove (Eds.). 2006a. *Disease Control Priorities in Developing Countries* (2nd ed.). New York: The World Bank and Oxford University Press.

Jamison, Dean T., J. G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, Prabhat Jha, Anne Mills, and Philip Musgrove (Eds.). 2006b. *Priorities in Health*. Washington, DC: World Bank.

Jepsen, Christopher, and S. Rivkin. 2009. "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size." *Journal of Human Resources* 44(1): 223-250.

Jimenez, Emmanuel, and H. A. Patrinos. 2008. "Can Cost-Benefit Analysis Guide Education Policy in Developing Countries." Policy Research Working Paper 4568. Washington, DC: World Bank.

Jimenez, Emmanuel, and Y. Sawada. 1999. "Do Community-Managed Schools Work? An Evaluation of El Salvador's EDUCO Program." *World Bank Economic Review 13*(3): 415-441.

JPAL. 2011. "Child Diarrhea." Downloaded November 30, 2011 from www.povertyactionlab.org/policy-lessons/health/child-diarrhea

Kremer, Michael, S. Moulin, and R. Namunyu. 2003. "Decentralization: A Cautionary Tale," unpublished working paper, Harvard University.

Kremer, Michael, E. Miguel, and R. Thornton. 2004. "Incentives to Learn." NBER Working Paper 10971. National Bureau of Economic Research.

Kremer, Michael, E. Miguel, and R. Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics* 91(3): 437-456.

Laxminarayan, Ramanan, J. Chow, and S. A. Shahid-Salles. 2006. "Intervention Cost-Effectiveness: Overview of Main Message." In Jamison, Dean T., Joel G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, Prabhat Jha, Anne Mills, and Philip Musgrove, (Eds.), *Disease Control Priorities in Developing Countries* (2nd ed.). New York: The World Bank and Oxford University Press.

Lee, David S., and T. Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48(2): 281-355.

Levin, Henry M., D. Catlin, and A. Elson. 2007. "Costs of Implementing Adolescent Literacy Programs." In Donald D. Deshler, Annemarie Sullivan Palincsar, Gina Biancarosa, and Marnie Nair, (Eds.), Informed Choices for Struggling Adolescent Readers. International Reading Association.

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

Levin, Henry M., and P. J. McEwan. 2001. *Cost-Effectiveness Analysis: Methods and Applications* (2nd ed.). Thousand Oaks, CA: Sage.

Levin, Henry M. and P. J. McEwan, eds. 2002. *Cost-Effectiveness and Educational Policy: 2002 Yearbook of the American Education Finance Association.* Larchmont, NY: Eye on Education.

Levine, Phillip B., and D. J. Zimmerman (Eds.). 2010. *Targeting Investments in Children: Fighting Poverty When Resources are Limited.* Chicago: University of Chicago Press.

Lockheed, Marlaine E., and E. Hanushek. 1988. "Improving Educational Efficiency in Developing Countries: What Do We Know?" *Compare* 18(1): 21-37.

Loeb, Susanna, and P. J. McEwan. 2006. "An Economic Approach to Education Policy Implementation." In Meredith I. Honig, (Ed.), *New Directions in Education Policy Implementation: Confronting Complexity.* State University of New York Press.

Loeb, Susanna, and P. J. McEwan. 2010. "Education Reforms." In Phillip B. Levine and David J. Zimmerman (Eds.), *Targeting Investments in Children: Fighting Poverty When Resources are Limited.* Chicago: University of Chicago Press.

McEwan, Patrick J. 2000. Constraints to Implementing Educational Innovations: The Case of Multigrade Schools." *International Review of Education* 46(1/2): 31-48.

_____. 2008. "Quantitative Research Methods in Education Finance and Policy." In Helen F. Ladd and Edward B. Fiske, eds., *Handbook of Research in Education Finance and Policy.* New York: Routledge.

_____. 2010. "The Impact of School Meals on Education Outcomes: Discontinuity Evidence from Chile." Wellesley College, unpublished manuscript.

Miguel, Edward, and M. Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1): 159-217.

Muennig, Peter. 2008. *Cost-Effectiveness Analyses in Health: A Practical Approach* (2nd ed.).

San Francisco: Jossey-Bass.

Murnane, Richard J., and J. B. Willett. 2011. *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* Oxford: Oxford University Press.

Muralidharan, Karthik, and V. Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy 119*(1): 39-77.

Musgrove, Philip, and J. Fox-Rushby. 2006. "Cost-Effectiveness Analysis for Priority Setting." In Dean T. Jamison, Joel G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, Prabhat Jha, Anne Mills, and Philip Musgrove (Eds.), *Disease Control Priorities in Developing Countries* (2nd ed.). New York: The World Bank and Oxford University Press.

Orr, L. L. 1999. *Social Experiments.* Thousand Oaks, CA: Sage.

Patrinos, H. A., and G. Psacharopoulos. 2010. "Returns to Education in Developing Countries." In Dominic Brewer and Patrick J. McEwan, (Eds.), *Economics of Education.* Amsterdam: Elsevier.

Ravallion, M. 2001. "The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation." *World Bank Economic Review 15*(1): 115-140.

_____. 2005. "Evaluating Anti-poverty Programs." Policy Research Working Paper 3625. Washington, DC: World Bank.

Rawlings, Laura B., and G. M. Rubio. 2005. "Evaluating the Impact of Conditional Cash Transfer Programs." *World Bank Research Observer* 20(1): 29-55.

Rosenthal, Robert. 1994. "Parametric Measures of Effect Size." In Harris Cooper and Larry V. Hedges, eds., *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Savedoff, William D., R. Levine, and N. Birdsall. 2006. *When Will We Ever Learn? Improving Lives Through Impact Evaluation.* Report of the Evaluation Gap Working Group. Washington, DC: Center for Global Development.

Schanzenbach, Diane Whitmore. 2007. "What Have Researchers Learned from Project STAR?" In Tom Loveless and Frederick M. Hess (Eds.), *Brookings Papers on Education Policy.* Washington, DC: Brookings Institution Press.

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74: 199-250.

Shadish, William R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin Company.

Skoufias, Emmanuel. 2005. "Progresa and Its Impacts on the Welfare of Rural Households in Mexico." IFPRI Research Report 139. Washington, DC: International Food Policy Research Institute.

Tan, J.-P., J. Lane, and P. Coustere. 1997. "Putting Inputs to Work in Elementary Schools: What Can Be Done in the Philippines?" *Economic Development and Cultural Change* 45(4): 857-879.

Tan-Torres Edejer, T., R. Baltussen, T. Adam, R. Hutubessy, A. Acharya, D. B. Evans, and C. J. L. Murray. 2003. *Making Choices in Health: WHO Guide to Cost-Effectiveness Analysis.* Geneva: World Health Organization.

Urquiola, Miguel S. 2006. "Identifying Class Size Effects in Developing Countries: Evidence from Rural Schools in Bolivia." *Review of Economics and Statistics* 88(1).

Vermeersch, Christel, and M. Kremer. 2005. "School Meals, Educational Achievement and School Competition." Policy Research Working Paper 3523. Washington, DC: World Bank.

What Works Clearinghouse (WWC). 2008. *Procedures and Standards Handbook (Version 2.0).* Washington, DC: Institute for Education Sciences, WWC. Downloaded June 29, 2011 from http://ies.ed.gov/ncee/wwc/references/idocviewer/ Doc.aspx?docId=19&tocId=5.

World Bank. 1993. *World Development Report: Investing in Health.* New York: Oxford University Press.

_____. 1997. *Primary Education in India.* Washington, DC: World Bank.

_____. 2008. *Global Purchasing Power Parities and Real Expenditures: 2005 International Comparison Program.* Washington, DC: World Bank.

_____. 2010. *Cost-Benefit Analysis in World Bank Projects.* Washington, DC: World Bank, Independent Evaluation Group.

Zwane, Alix Peterson, and M. Kremer. 2007. "What Works in Fighting Diarrheal Diseases in Developing Countries?" NBER Working Paper 12987. National Bureau of Economic Research.