

Cost-effectiveness analysis of education and health interventions in developing countries

Patrick J. McEwan*

Department of Economics, Wellesley College, 106 Central Street, Wellesley, MA 02481, USA

High-quality impact evaluations, including randomised 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-utilised 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, emphasising education and health interventions, and discusses how the methods are currently applied in the literature.

Keywords: cost-effectiveness analysis; education; health; impact evaluation

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 2004, Duflo and Kremer 2005, Savedoff *et al.* 2006). Fortunately, the quantity of impact evaluations in education and health grew rapidly, and they increasingly applied high-quality experimental research designs.¹ The growing number of impact evaluations has facilitated reviews of the most effective interventions in education and health.²

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 (Urquiola 2006, Schanzenbach 2007). 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).

There is broad consensus that cost-effectiveness analysis (CEA) and cost-benefit analysis (CBA) are necessary to inform resource allocation decisions, and a substantial literature codifies methods to conduct cost analyses of education and health interventions (Gold *et al.*

*Email: pmcewan@wellesley.edu

1996, Levin and McEwan 2001, Drummond *et al.* 2005). However, some evidence suggests that the methods are under-utilised or imperfectly applied. For example, the World Bank mandates that projects conduct economic analysis to determine whether project benefits outweigh costs (World Bank 1994). If benefits cannot be calculated in monetary terms, then CEA is an acceptable alternative method. A comprehensive review of projects shows that the use of CBA has declined since the 1970s, both at the time of project appraisal and project closure (World Bank 2010). Across multiple sectors, education and health projects are, by far, among the least likely to report results of a CBA at project appraisal.³ Nor does it appear that CEA is used instead: of 24 projects that purported to conduct a CEA, only one actually did (World Bank 2010).⁴ It is reasonable to conclude that *ex ante* cost-effectiveness in education and health projects is relatively rare and, when applied, is often misconstrued as a simple cost analysis or as a cost-benefit method capable of judging the potential worth of a single intervention.

The academic literature on CEA justifies more optimistic conclusions. In developing countries, the cost-effectiveness literature has grown in health policy, particularly with the most recent *Disease Control Priorities in Developing Countries* and its predecessors (Jamison *et al.*, 1993, 2006a, 2006b). Many recent randomised experiments in health and education have often been accompanied by a CEA of interventions that pursue similar objectives (for example, Miguel and Kremer 2004, Banerjee *et al.* 2007). With the recent growth in the use of rigorous impact evaluation methods, it is a propitious moment to review the methods and recent applications of CEA, with a particular 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 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 is 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 CEA as it currently being conducted in developing countries, ranging from single-study CEA embedded within a randomised 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 (CERs), and the appropriate steps to follow in conducting *ex post* and *ex ante* CEA. Section 7 concludes.

2. Framework and definitions

2.1. Types of cost analysis

Cost analysis falls into two broad categories: CBA and CEA. A third approach, cost-utility analysis, 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 labour-market outcomes. The intervention may be a small-scale programme (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-intervention 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 labour. 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 incremental benefits or as 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.

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, one needs to estimate the net present value (NPV) of an intervention:

$$\text{NPV} = \sum_{t=0}^n \frac{B_t}{(1+r)^t} - \sum_{t=0}^n \frac{C_t}{(1+r)^t},$$

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 who discounts future benefits (or costs) less heavily.⁷

The NPV has a convenient interpretation as the absolute desirability of the intervention (whether positive or negative). Its magnitude can also be compared with 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 enrolment, 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 enrolment or attendance, health incidents averted (for example, respiratory or diarrhoeal 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 CER is C / E . This represents the incremental cost per unit of incremental effect (that is, the cost of enrolling another child in school, or the cost of saving an additional life). Authors also report effectiveness–cost ratios (E / C , or units of incremental effect per unit of incremental cost), although this 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 similar terms, as the incremental cost per student subjected to the intervention.

Table 1 summarises a CEA of Kenyan education interventions to improve child test scores, adapted from Kremer *et al.* (2004). The authors conducted an experimental impact evaluation of a programme that provided merit scholarships for adolescent girls who scored well on examinations. 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 with 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?

In Table 1, the authors calculated CERs for other interventions, using other Kenyan experimental evaluations, including a teacher incentive programme, textbooks and flip-chart provision, and school-based deworming. The effect of some interventions could not be 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 programme impact, do not provide unambiguous guidance about resource allocation. For example, Table 1 indicates that the scholarship programme as implemented in the Busia district is relatively more effective and cost-effective than the larger programme. It also happens to be more cost-effective than teacher incentives. The result highlights the importance of: assessing the sensitivity of CERs to alternate assumptions about costs and effects; and carefully considering the external validity of estimated CERs, especially when generalising 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 (Miguel and Kremer 2004, Dhaliwal *et al.* 2011). 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.

Table 1. Cost-effectiveness ratios of education interventions in Kenya.

Intervention	Effect	Cost	Cost-effectiveness ratio	
	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 programme				
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 programme	≈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 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 C/U reflects the incremental cost per unit of incremental utility. Cost-utility analysis 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 two years (from eight to 10), measured on the x axis. It also improves quality of each year lived, measured on the y axis, where one indicates perfect health, zero 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).

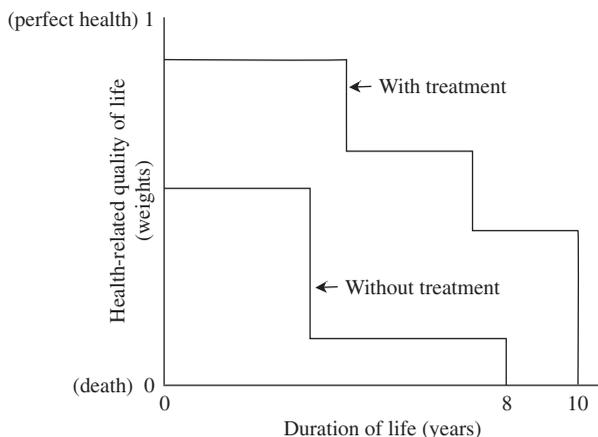


Figure 1. An illustration of quality-adjusted life years (QALYs).
Source: Adapted from Drummond *et al.* (2005).

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 (for example, 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 summarised 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 and \$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 counselling 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 generalisability 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: governments, including agencies directly and indirectly involved in implementing the intervention; non-governmental organisations or private firms; social sector clients, such as students, patients, and their families; and non-clients

who nonetheless 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 labour. 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 (for example, 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 neighbouring non-intervention schools. In the United States, 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 monetised, 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 effects received by clients.

2.6. *Ex post CEA 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 organisation, the CEA is often *ex ante* (World Bank 2010). 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 important considerations when gathering evidence on effects and costs, respectively. Section 6 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.

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 (Shadish *et al.* 2002, McEwan 2008). 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).¹¹ 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 randomised experiment or a high-quality quasi-experiment.

An estimate of effectiveness is externally valid when it can be generalised to different versions of the intervention, to different samples of subjects, and to different policy contexts. For example, conditional cash transfer programmes have been implemented in many countries (Fiszbein and Schady 2009). It is not always certain whether estimated effects in one programme can be generalised 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 Conditional Cash Transfers (CCTs) has shown. More often, especially for little-researched interventions, judgements about external validity are not clear-cut and are informed by commonsense and theory. The issue of external validity in CEA is revisited in Section 6.

3.2. *Methods of estimating effectiveness*

The literature in impact evaluation focuses overwhelmingly on the importance of improving internal validity.¹² 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: evaluation design and statistical controls.

3.2.1. Evaluation design. Sound evaluation design, especially randomised experiments and good quasi-experiments, are the best means of ruling out selection as a threat to internal validity. In the classic randomised 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.¹³ For example, the Honduran PRAF programme awarded cash transfers to eligible families living in entire municipalities that were randomly assigned to participate in the programme. 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 in education and health outcomes could be causally attributed to the treatment. In quasi-experiments, the treatment assignment may contain elements of randomness or purposeful assignment, but some might be due to the choices of students, parents, or administrators. When greater control is exerted, then causal results often possess greater internal validity.¹⁴

3.2.2. Statistical controls. Researchers use statistical methods – alone or in concert with evaluation design – to control for observed pre-existing differences between treatment and control groups. 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.¹⁵ 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 randomised 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 (for example, 'mathematics achievement') but different tests and measurement scales. Thus, a point gained on one test cannot be meaningfully compared with 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 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 from tests that 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 modelling (often using a secondary dataset). For example, JPAL (2011) compares the cost-effectiveness of multiple interventions in reducing the incidence of child diarrhoea, 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 diarrhoea.

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 modelling 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 Progreso programme 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 United States, 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 one-standard-deviation increase in test scores yields a 10 per cent 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.¹⁶

The prior literature reveals a trade-off between internal validity and the potential richness of outcome measures in a CEA. Randomised 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 modelling. The latter exercise should also be accompanied by additional sensitivity analysis, especially on the assumed or estimated parameters.

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.¹⁷ 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 emphasise that the objective is to estimate incremental costs incurred over the duration of the intervention that was actually

evaluated. In recent randomised 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 programme 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 utilises 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 labour, a shadow price must be estimated.

4.2. Specification of ingredients

The specification of ingredients is often facilitated by dividing ingredients into categories including: personnel, facilities, equipment and materials, other programme inputs, and 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 (for example, Fiedler *et al.* 2008).

4.2.1. 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 specialised skills required for the positions. Time commitments refer to the amount of time 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.¹⁸

4.2.2. 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.

4.2.3. 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.

4.2.4. 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.

4.2.5. Required client inputs. This category of ingredients includes any contributions that are 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 programme. 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 programme decentralised 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 10 per cent 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 per cent 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: through document review; through interviews with personnel or observations of the intervention; and through empirical analysis. An essential starting point is the examination of programme documents. These documents include general descriptions of the programme prepared by programme staff or outsiders, budgets and expenditure statements, web sites, and reports by previous evaluators of the programme.

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 whether the prescribed ingredients are actually being used. If the programme designer mentioned that students should have individual workbooks, is it the case that all students in the class have workbooks? If programme 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 programme implementation is a common occurrence, even in 'best-practice' education programmes (McEwan 2000, Loeb and McEwan 2006). In the United States, Levin *et al.* (2007) document that costs of adolescent literacy programmes can vary widely across sites due to variable programme implementation, even though each site implements a nominally homogeneous 'programme'. Their analysis illustrates the potential biases of cost estimates that rely exclusively on back-of-the-envelope calculations using programme documents.

A third source of information is empirical analysis of intervention data, ideally collected in the impact evaluation itself. Randomised 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: they are often collected for both treatment and control groups, allowing for a more careful identification of incremental ingredients consumed; and 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 programme 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 organisations 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 amortise the total value of the facility over its remaining useful life, estimating an annual cost of the facility. The annualised 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 30-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 (for example, laboratory equipment or textbooks) have a usable life of more than one year, those costs can also be annualised. For example, a CEA 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 annualise costs (Harbison and Hanushek 1992). The concepts and formulas for annualisation are discussed in Levin and McEwan (2001) and Drummond *et al.* (2005).

4.5. Adjusting costs for inflation, time-value, and currency

4.5.1. *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 summarised in a price index such as the consumer price index or the gross domestic product 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) × \$120].

4.5.2. *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 following 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, and an extensive theoretical and empirical literature informs its choice (Boardman *et al.* 2011).

From an applied standpoint, two issues are important. First, the social discount rate is often mandated by the organisation sponsoring the CEA. For example, the Inter-American Development Bank applies a consistent rate of 12%. Dhaliwal *et al.* (2011) summarise 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.

4.5.3. *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 US dollars). However, poorer countries tend to have relatively lower prices for non-traded goods and services, and local currency actually has more purchasing power than it would in global markets. Purchasing power parity exchange 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 purchasing power parity exchange rates.

5. Who does CEA in developing countries?

5.1. CEA embedded in non-experimental impact evaluations

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

5.1.1. Rate-of-return studies. The first approach yielded hundreds of estimates of the rates of return to various levels of schooling (Jimenez and Patrinos 2008, Patrinos and Psacharopoulos 2010). 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 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.¹⁹ 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 per cent.

5.1.2. Education production function studies. Education economists frequently 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 education production functions studies combined cost data with coefficient estimates to conduct a full CEA.²⁰ 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, annualising 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.

5.2. CEA embedded in experimental impact evaluations

Randomised experiments have improved greatly on the internal validity of non-experimental regression studies and they often report a 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 summarised in Table 1.²¹ 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 favourably with other options. Finally, they conducted a partial CBA by using a rate of return 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 rate-of-return estimates obtained from the literature or supplemental analysis of household survey data.

Second, even randomised 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 programme 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.

5.3. CEA league tables

5.3.1. 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 emphasised non-experimental evaluations. More recently, Evans and Ghosh (2008) utilise a broad base of experimental and non-experimental studies that attempt to measure impacts on school enrolment, school participation, and test scores. The authors standardise effectiveness measures when appropriate (for example, converting test score effects to effect sizes). Costs are converted to annualised costs when appropriate, or a present value when an intervention is evaluated over several years. Finally, they convert costs to 1997 US 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 US 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.²² The assumptions of the CEA are described by Dhaliwal *et al.* (2011).

5.3.2. *Health interventions.* International league tables for health interventions commonly adopt the DALY as a measure of effectiveness. The most ambitious effort to date is the second *Disease Control Priorities in Developing Countries* (Jamison *et al.* 2006a, 2006b), which commissioned reviews of interventions' effectiveness and costs and summarised 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) summarised 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, and so on. Finally, the Poverty Action Lab reports basic league tables for the cost-effectiveness of increasing healthcare provider attendance and reducing child diarrhoea, 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.²³ 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 organisations. Whatever the standard applied within an organisation, it is almost certain that a CEA will compare its results with 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 labour 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 per cent 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.

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 generalised outside the context of the initial evaluation.

6.2.1. Interventions. The highest-quality impact evaluations are often randomised 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 (although the limited evidence from Progresa suggests otherwise; see Fiszbein and Schady 2009).

A classic example in the United States 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, state-wide implementation of class size reduction in California, which immediately created 25,000 new teaching positions. In competitive labour 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.

6.2.2. Subjects. Impact evaluations typically report average treatment effects that can be generalised 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 generalised 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 generalisability. 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 emphasised 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 generalisable to lower-income populations of Latin America.

6.2.3. *Settings.* Finally the setting of an intervention can have great bearing on whether effects and costs can be generalised. Baseline conditions often vary substantially from country to country. In Mexico's Progreso evaluation, the net enrolment rate is over 90 per cent among young, primary-aged children, and relatively small enrolment effects were observed for that group. In a Honduran evaluation, where net enrolments are much lower, the effects were larger.²⁴

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 of 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 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 randomised experiments in its CEAs (Dhaliwal *et al.* 2011). The US What Works Clearinghouse, which reviews US studies of education interventions, has more ecumenical guidelines (What Works Clearinghouse 2008). While they also prize randomised experiments, What Works Clearinghouse standards concede that some quasi-experiments can meet standards of high-quality evidence, especially regression-discontinuity designs. These standards also allow evidence from randomised 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 judgements 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, in so far as they are available. In the best circumstances they are reported in the evaluation. More probable is that they are located in the unpublished grey literature of working papers or commissioned reports. Finally, it is possible that no cost estimate exists in which case researchers might attempt to either reconstruct an ingredients-based estimate using documents, interviews, and secondary data analysis, or use a cost estimate of a comparable intervention from another setting (as in Miguel and Kremer 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 labour, unless the analysis explicitly adopts a government or funder perspective. Second, one should take care to express the incremental cost estimate in units comparable with 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 five 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* five 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-utilised 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 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 systematise methods for calculating and comparing incremental costs, in much the same way that impact evaluators have done for the measurement of effectiveness.

Finally, it bears emphasis that impact evaluation and accompanying CBA or CEA are not the only ingredients to policy decisions. Among other factors, the policy decision must weigh the financial sustainability of the policy and political constraints to its effective implementation and scale-up. This paper has argued that good cost analysis is a very necessary, albeit insufficient condition for sensible decisions about resource allocation in health and education.

Acknowledgements

The author is grateful to Henry Levin for prior collaborations that greatly informed his thinking. Alaka Holla, Francisco Mejia, Paul Winters, and anonymous referees provided helpful comments.

Notes

1. Prominent examples included the experimental evaluations of conditional cash transfers in Mexico and school-based deworming treatments in Kenya (Miguel and Kremer 2004, Schultz 2004, Skoufias 2005).
2. See, for example, Rawlings and Rubio (2005) 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) summarise impact evaluations from broad array of evaluation literature in health, education, and other sectors.
3. Only 1 per cent of education and health projects reported an economic rate of return at project entry (World Bank 2010).
4. In the US education literature, Levin and McEwan (2001, 2002) were similarly pessimistic about quantity and quality of CEA. Clune (2002) showed that the term cost-effectiveness is predominantly rhetorical in US education literature, and is rarely accompanied by a concrete empirical analysis, similar to the World Bank (2010) review.
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 (2011), Imbens and Wooldridge (2009), Murnane and Willett (2011), and McEwan (2008).
7. Related measures include the internal rate of return (the 'break-even' value of r that equalises 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.
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.
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-Rushby 2006). See Drummond *et al.* (2005, p. 187) and Musgrove and Fox-Rushby (2006) for a further discussion of DALYs and comparison with QALYs.
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 Progresá programme were almost 5 per cent of the total transfer amounts.

11. Shadish *et al.* (2002) describe the history of the counterfactual reasoning. Statisticians formalised the framework, especially Donald Rubin (Holland 1986), in a framework that has been adopted within impact evaluation (for example, Ravallion 2005).
12. 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 (2011), and Shadish *et al.* (2002).
13. A coin flip or similar mechanism is only the simplest approach to designing randomised 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 randomised experiments, see Orr (1999) and Duflo *et al.* (2006).
14. One of the most credible quasi-experimental methods is the regression-discontinuity design, in which assignment is determined by a continuous assignment variable and a specified cutoff; see especially Lee and Lemieux (2010) and DiNardo and Lee (2011). Lee and Lemieux (2010) interpret the regression-discontinuity design as a 'local' randomised experiment, since the precise location of subjects on either side of the cutoff is partly due to random noise in the assignment variable. Other quasi-experimental methods (for example, 'natural' experiments occasioned by variation in an instrumental variable) can vary substantially in their internal validity, depending on the empirical context.
15. 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 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 observably 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 (Imbens and Wooldridge 2009).
16. Muralidharan and Sundararaman (2011) report an exception from India.
17. For similar descriptions in education and health, respectively, see Levin and McEwan (2001, Chapters 3–5) and Drummond *et al.* (2005, Chapter 4).
18. 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.
19. Literature in developed countries focused extensively on methods to correct for selection bias, including twins studies and instrumental variables estimates (Gunderson and Oreopoulos 2010). The developing-country literature is surprisingly weak in this area, although Patrinos and Psachropoulos (2010) review some exceptions.
20. 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.
21. 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.
22. See <http://www.povertyactionlab.org/policy-lessons> [Accessed 30 November 2011].
23. See Boardman *et al.* (2011) and Levin and McEwan (2001) for further discussion.
24. See Galiani and McEwan (2011) and Fiszbein and Schady (2009, Table 5.1).

References

- Angrist, J.D. and Krueger, A.B., 1999. Empirical strategies in labor economics. *In*: O. Ashenfelter and D. Card, eds. *Handbook of labor economics*. Vol. 3A. Amsterdam: Elsevier, 1277–1366.
- Banerjee, A.V., *et al.*, 2005. *Remedying education: evidence from two randomized experiments in India*. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 11904.
- Banerjee, A.V., *et al.*, 2007. Remedying education: evidence from two randomized experiments in India. *Quarterly journal of economics*, 122 (3), 1235–1264.
- Bedi, A.S. and Marshall, J.H., 1999. School attendance and student achievement: evidence from rural Honduras. *Economic development and cultural change*, 47 (3), 657–682.
- Boardman, A.E., *et al.*, 2011. *Cost–benefit analysis: concepts and practice*. 4th ed. Boston: Prentice Hall.

- Bouillon, C.P. and Tejerina, L., 2006. *Do we know what works? A systematic review of impact evaluations of social programs in Latin America and the Caribbean*. Washington, DC: Inter-American Development Bank, Sustainable Development Department, Poverty and Inequality Unit.
- Caldés, N., Coady, D., and Maluccio, J.A., 2006. The cost of poverty alleviation transfer programs: a comparative analysis of three programs in Latin America. *World development*, 34 (5), 818–837.
- Canning, D., 2006. The economics of HIV/AIDS in low-income countries: the case for prevention. *Journal of economic perspectives*, 20 (3), 121–142.
- Clune, W.H., 2002. Methodological strength and policy usefulness of cost-effectiveness research. In: H.M. Levin and P.J. McEwan, eds. *Cost-effectiveness and educational policy: 2002 yearbook of the American Education Finance Association*. Larchmont, NY: Eye on Education, 55–68.
- Dhaliwal, I., et al., 2011. Comparative cost-effectiveness to inform policy in developing countries. Unpublished manuscript. Abdul Latif Jameel Poverty Action Lab, MIT.
- DiNardo, J. and Lee, D.S., 2010. Program evaluation and research designs. In: O. Ashenfelter and D. Card, eds. *Handbook of labor economics*. Vol. 4A. Amsterdam: Elsevier, 463–536.
- Drummond, M.F., et al., 2005. *Methods for the economic evaluation of health care programmes*. 3rd ed. Oxford: Oxford University Press.
- Duflo, E., 2001. School and labor market consequences of school construction in Indonesia: evidence from an unusual policy experiment. *American economic review*, 91 (4), 795–813.
- Duflo, E., 2004. Scaling up and evaluation. Paper prepared for the Annual World Bank Conference on Development Economics.
- Duflo, E. and Kremer, M., 2005. Use of randomization in the evaluation of development effectiveness. In: G.K. Pitman, O.N. Feinstein, and G.K. Ingram, eds. *Evaluating development effectiveness*. New Brunswick, NJ: Transaction Publishers, 205–231.
- Duflo, E., Glennerster, R., and Kremer, M., 2006. Using randomization in development economics research: a toolkit. Unpublished manuscript. MIT.
- Duncan, G.J., Ludwig, J., and Magnuson, K.A., 2010. Child development. In: P.B. Levine and D.J. Zimmerman, eds. *Targeting investments in children: fighting poverty when resources are limited*. Chicago: University of Chicago Press, 27–58.
- Evans, D. K. and Ghosh, A., 2008. *Prioritizing educational investments in children in the developing world*. Santa Monica, CA: RAND, Labor and Population, Working Paper WR-587.
- Fiedler, J.L., Villalobos, C.A., and De Mattos, A.C., 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, A. and Schady, N., 2009. *Conditional cash transfers: reducing present and future poverty*. Washington, DC: World Bank.
- Fuller, B., Hua, H., and Snyder, C.W., 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, S. and McEwan, P.J., 2011. The heterogeneous impact of conditional cash transfers. Unpublished manuscript. Washington University and Wellesley College.
- Glewwe, P., 1999. *The economics of school quality investments in developing countries: an empirical study of Ghana*. London: St. Martin's.
- Glewwe, P. and Lambert, S., 2010. Education production functions: evidence from developing countries. In: D. Brewer and P.J. McEwan, eds. *Economics of education*. Amsterdam: Elsevier, 137–147.
- Glewwe, P., Kremer, M., and Moulin, S., 1997. Textbooks and test scores: evidence from a prospective evaluation in Kenya. Unpublished manuscript, University of Minnesota.
- Glewwe, P., Ilias, N., and Kremer, M., 2003. *Teacher incentives*. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 9671.
- Glewwe, P., et al., 2004. Retrospective vs. prospective analysis of school inputs: the case of flip charts in Kenya. *Journal of development economics*, 74 (1), 251–268.
- Gold, M.R., et al., eds, 1996. *Cost-effectiveness in health and medicine*. New York: Oxford University Press.
- Gunderson M. and Oreopoulos, P., 2010. Returns to education in developed countries. In: D. Brewer and P.J. McEwan, eds. *Economics of education*. Amsterdam: Elsevier, 37–43.
- Harbison, R.W. and Hanushek, E.A., 1992. *Educational performance of the poor: lessons from rural northeast Brazil*. Oxford: Oxford University Press.
- Holla, A. and Kremer, M., 2009. *Pricing and access: lessons from randomized evaluations in education and health*. Washington, DC: Center for Global Development, Working Paper 158.

- Holland, P.W., 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396), 945–960.
- Imbens, G.W. and Wooldridge, J.M., 2009. Recent development in the econometrics of program evaluation. *Journal of economic literature*, 47 (1), 5–86.
- Jamison, D.T., et al., ed., 1993. *Disease control priorities in developing countries*. New York: Oxford University Press.
- Jamison, D.T., et al., eds, 2006a. *Disease control priorities in developing countries*. 2nd ed. New York: The World Bank and Oxford University Press.
- Jamison, D.T., et al., eds, 2006b. *Priorities in health*. Washington, DC: World Bank.
- Jepsen, C. and Rivkin, S., 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, E. and Patrinos, H.A., 2008. *Can cost-benefit analysis guide education policy in developing countries*. Washington, DC: World Bank, Policy Research Working Paper 4568.
- Jimenez, E. and Sawada, Y., 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. Available from: www.povertyactionlab.org/policy-lessons/health/child-diarrhea [Accessed 30 November 2011].
- Kremer, M., Moulin, S., and Namunyu, R. 2003. Decentralization: a cautionary tale. Unpublished working paper. Harvard University.
- Kremer, M., Miguel, E., and Thornton, R., 2004. *Incentives to learn*. Cambridge, MA: National Bureau of Economic Research, Working Paper No. 10971.
- Kremer, M., Miguel, E., and Thornton, R., 2009. Incentives to learn. *Review of economics and statistics*, 91 (3), 437–456.
- Laxminarayan, R., Chow, J., and Shahid-Salles, S.A., 2006. Intervention cost-effectiveness: overview of main message. In: D.T. Jamison et al., eds. *Disease control priorities in developing countries*. 2nd ed. New York: The World Bank and Oxford University Press, 35–86.
- Lee, D.S. and Lemieux, T., 2010. Regression discontinuity designs in economics. *Journal of economic literature*, 48 (2), 281–355.
- Levin, H.M. and McEwan, P.J., 2001. *Cost-effectiveness analysis: methods and applications*. 2nd ed. Thousand Oaks, CA: Sage.
- Levin, H.M. and McEwan, P.J., eds, 2002. *Cost-effectiveness and educational policy: 2002 yearbook of the American Education Finance Association*. Larchmont, NY: Eye on Education.
- Levin, H.M., Glass, G.V., and Meister, G.R., 1987. Cost-effectiveness of computer-assisted instruction. *Evaluation review*, 11 (1), 50–72.
- Levin, H.M., Catlin, D., and Elson, A., 2007. Costs of implementing adolescent literacy programs. In: D.D. Deshler et al., eds. *Informed choices for struggling adolescent readers*. International Reading Association, 61–91.
- Levine, P.B. and Zimmerman, D.J., eds, 2010. *Targeting investments in children: fighting poverty when resources are limited*. Chicago: University of Chicago Press.
- Lockheed, M.E. and Hanushek, E., 1988. Improving educational efficiency in developing countries: what do we know? *Compare*, 18 (1), 21–37.
- Loeb, S. and McEwan, P.J., 2006. An economic approach to education policy implementation. In: M.I. Honig, ed. *New directions in education policy implementation: confronting complexity*. New York: State University of New York Press, 169–186.
- Loeb, S. and McEwan, P.J., 2010. Education reforms. In: P.B. Levine and D.J. Zimmerman, eds. *Targeting investments in children: fighting poverty when resources are limited*. Chicago: University of Chicago Press, 145–178.
- McEwan, P.J., 2000. Constraints to implementing educational innovations: the case of multigrade schools. *International review of education*, 46 (1/2), 31–48.
- McEwan, P.J., 2008. Quantitative research methods in education finance and policy. In: H.F. Ladd and E.B. Fiske, eds. *Handbook of research in education finance and policy*. New York: Routledge, 87–104.
- McEwan, P.J., 2010. The impact of school meals on education outcomes: discontinuity evidence from Chile. Unpublished manuscript. Wellesley College.
- Miguel, E. and Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72 (1), 159–217.
- Muennig, P., 2008. *Cost-effectiveness analyses in health: a practical approach*. 2nd ed. San Francisco: Jossey-Bass.

- Muralidharan, K. and Sundararaman, V., 2011. Teacher performance pay: experimental evidence from India. *Journal of political economy*, 119 (1), 39–77.
- Murnane, R.J. and Willett, J.B., 2011. *Methods matter: improving causal inference in educational and social science research*. Oxford: Oxford University Press.
- Musgrove, P. and Fox-Rushby, J., 2006. Cost-effectiveness analysis for priority setting. In: D.T. Jamison *et al.*, eds. *Disease control priorities in developing countries*. 2nd ed. New York: The World Bank and Oxford University Press, 271–286.
- Orr, L.L., 1999. *Social experiments*. Thousand Oaks, CA: Sage.
- Patrinos, H.A. and Psacharopoulos, G., 2010. Returns to education in developing countries. In: D. Brewer and P.J. McEwan, eds. *Economics of education*. Amsterdam: Elsevier, 44–51.
- Ravallion, M., 2001. The mystery of the vanishing benefits: an introduction to impact evaluation. *World Bank economic review*, 15 (1), 115–140.
- Ravallion, M., 2005. *Evaluating anti-poverty programs*. Washington, DC: World Bank, Policy Research Working Paper No. 3625.
- Rawlings, L.B. and Rubio, G.M., 2005. Evaluating the impact of conditional cash transfer programs. *World Bank research observer*, 20 (1), 29–55.
- Rosenthal, R., 1994. Parametric measures of effect size. In: H. Cooper and L.V. Hedges, eds. *The handbook of research synthesis*. New York: Russell Sage Foundation, 231–244.
- Savedoff, W.D., Levine, R., and Birdsall, N., 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, D.W., 2007. What have researchers learned from Project STAR? In: T. Loveless and F.M. Hess, eds. *Brookings papers on education policy*. Washington, DC: Brookings Institution Press, 205–228.
- Schultz, T.P., 2004. School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of development economics*, 74, 199–250.
- Shadish, W.R., Cook, T.D., and Campbell, D.T., 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.
- Skoufias, E., 2005. *PROGRESA and its impacts on the welfare of rural households in Mexico*. Washington, DC: International Food Policy Research Institute, Research Report 139.
- Tan, J.-P., Lane, J., and Coustere, P., 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., *et al.*, 2003. *Making choices in health: WHO guide to cost-effectiveness analysis*. Geneva: World Health Organization.
- Urquiola, M.S., 2006. Identifying class size effects in developing countries: evidence from rural schools in Bolivia. *Review of economics and statistics*, 88 (1), 171–177.
- Vermeersch, C. and Kremer, M., 2005. *School meals, educational achievement and school competition*. Washington, DC: World Bank, Policy Research Working Paper 3523.
- What Works Clearinghouse, 2008. *Procedures and standards handbook (version 2.0)*. Washington, DC: Institute for Education Sciences. Available from: <http://ies.ed.gov/ncee/wwc/references/idocviewer/Doc.aspx?docId=19&tocId=5> [Accessed 29 June 2011].
- World Bank, 1993. *World development report: investing in health*. New York: Oxford University Press.
- World Bank, 1994. Operational policy 10.04. Available from: <http://go.worldbank.org/HEENE0JRS0> [Accessed 7 December 2011].
- World Bank, 1997. *Primary education in India*. Washington, DC: World Bank.
- World Bank, 2008. *Global purchasing power parities and real expenditures: 2005 international comparison program*. Washington, DC: World Bank.
- World Bank, 2010. *Cost-benefit analysis in World Bank projects*. Washington, DC: World Bank, Independent Evaluation Group.
- Zwane, A.P. and Kremer, M., 2007. *What works in fighting diarrheal diseases in developing countries?* Cambridge, MA: National Bureau of Economic Research, Working Paper No. 12987.