Toward Policy-Relevant Benchmarks for Interpreting Effect Sizes: Combining Effects with Costs

Douglas N. Harris
University of Wisconsin at Madison

September 17, 2008

Abstract: The common reporting of effect sizes has been an important advance in education research in recent years. However, the benchmarks used to interpret the size of these effects—as small, medium, and large—do little to inform educational administration and policymaking because they do not account for program costs. I propose an approach to establishing cost-effectiveness benchmarks rooted in an explicit economics-based decision-making framework and assumptions about the decision-making context. To be considered "large," the ratio of effects-to-costs must be at least as large as the ratios of substitute interventions. Evidence related to class size, pre-kindergarten and other interventions is discussed to illustrate the calculation of the cost-effectiveness ratios, how the evidence can be used to develop benchmarks, and how the benchmarks can be useful for researchers and policymakers. The development of benchmarks is intended to encourage cost-effectiveness analysis as a standard part of policy analysis, thereby providing more evidence to increase the validity of the benchmarks, and, ultimately, improving policy decisions. Recent cost-effectiveness research in health care policy illustrates the potential value of cost-effectiveness benchmarks in education.

Author Information: Douglas N. Harris, Department of Educational Policy Studies, 213 Education Building, 1000 Bascom Mall, Madison, WI 53706. Phone: 608-262-4406; Email: dnharris3@wisc.edu.

Acknowledgements: The author wishes to thank David Grissmer, Thomas Luschei, and three anonymous reviewers for providing especially detailed and useful comments on various drafts. In addition, the author wishes to thank Geoffrey Borman, Adam Gamoran, Henry Levin, and participants in a seminar at the University of Wisconsin at Madison, Interdisciplinary Training Program. The author is responsible for all remaining errors.

Introduction

Nearly all empirical educational research focuses on the effects of interventions, programs, and policies—and ignores costs (Levin, 1991; Rice, 2002). Consideration of cost is especially rare in education compared with other areas of public policy (Monk and King, 1993; Rice, 2002). One reason is that economists generally have shown little interest in applying these economics-based techniques to education. Another is that the norms of practice in educational research do not encourage consideration of costs. As a result, there are few studies that consider costs and therefore little basis of comparison for researchers who might wish to consider them. This has created something of a "catch-22." There have been few cost analyses because there has been no basis of comparison and no basis of comparison because there have been so few cost analyses. This study attempts to break the catch-22 by: (1) highlighting the importance of integrating the usual evidence about effects with cost analysis; and (2) outlining a specific approach to developing cost-effectiveness based benchmarks to help scholars and policymakers interpret educational policy research.

To see why costs are important, consider how empirical educational research is typically conducted. The first step is to estimate an effect of an intervention (program, policy, etc.) on an outcome such as student achievement. These effects are then converted to effect sizes so that a one unit change in the dependent variable represents one standard deviation. Finally, it is common to interpret effect sizes in terms of the benchmarks identified by Cohen (1988): an effect size is "small" if it is near 0.2, medium if it is near 0.5, and large if it is near or larger than 0.8. Cohen himself emphasizes that these benchmarks are somewhat arbitrary and should not be strictly applied, but this has not stopped just the sort of simple applications that he warns against.[1]

In addition to being arbitrary, these effect size benchmarks tell only half the story. Effects are obviously important, but because policymakers also face constraints on their budgets and other resources, they cannot do their best to improve outcomes without knowing which policies will contribute the most *relative to their costs*. Take the case of pre-kindergarten. As I discuss below, some versions of this policy may increase student test scores by as much as 0.5 standard deviations, the same as the "medium" benchmark established by Cohen. But it is also extremely costly at more than $18,000 per pupil. This does not mean that pre-kindergarten should not be adopted—on the contrary, I show later that this policy compares favorably with many others even after accounting for these high costs—but it obviously would be better if this same effect size could be obtained with lower costs through alternative interventions. In short, there is no reason to expect any relationship between the Cohen benchmarks and the decision-making benchmarks that are appropriate for policymaking. A statistician might consider an effect size of 0.5 to be of medium size, but a policymaker might see it as small or large.

In this study, I propose an approach to establishing cost-based benchmarks for interpreting effect sizes and outline the necessary assumptions. The approach is comprised of two main parts: an economics-based decision framework and a set of assumptions about the decision-making context. The decision *framework* is based on the simple and intuitive economics notion that a policy should be adopted only if there is no other way to create the same effect at a lower cost. The most important parts of the decision *context* are arguably the decision-makers themselves. For example, decision-makers must be trying to maximize some specific outcomes(s) that are covered by the benchmarks. They must also understand the breadth of policy options at their disposal and the "substitutability" of the options on

which the benchmarks are based. Some decision-makers, especially those at the state and federal levels, have the authority to adopt very different types of policies (e.g., class size reduction and whole school reform), which I call "far substitutes." Other decision-makers only have responsibility over a narrow range of options (e.g., different types of teacher professional development), or "near substitutes." Given the assumptions of the decision context and the framework, an intervention with a cost-effectiveness ratio as least as large as the ratios of the relevant substitutes is considered "large" and policymakers should therefore give it serious consideration. Conversely, interventions with "small" cost-effectiveness ratios are less cost-effective compared with the relevant substitutes and policymakers therefore should be cautious about adopting them.

Establishing cost-based benchmarks is, as a practical matter, about calculating cost-effectiveness ratios and comparing them across interventions. Lockheed and Hanushek (1988) and Schiefelbain, Wolff, & Scheifelbain (1999) compare the cost-effectiveness of interventions in developing countries, but do not explicitly consider issues of comparability, how the use of the evidence depends on the decision-making context, or how the evidence might be used to create benchmarks. Rice (1997) introduces some of the issues that arise in comparing cost-effectiveness ratios, focusing on how the use of cost-effectiveness analysis might be expanded through a framework that encourages educational practitioners and program evaluators to consider site-specific cost information. In contrast, the benchmarking approach proposed here is intended to increase the use of cost analysis among researchers—still in an effort to inform local and other decisions, but in a way that fits the focus of researchers on theory and generalizability.[2] Rice's approach and the one proposed here are

therefore different in substance, but still complementary in their larger aim of expanding the use of cost-effectiveness analysis as a tool for policymaking.

Compared with education, the use of cost analysis is much more common in health research. Researchers in that field have also developed benchmarks of the sort proposed here. One prominent report, using years of life saved as the health outcome of interest, suggests a benchmark of $100 per life year for interventions in developing countries (Laxminarayan, Chow, & Shahid–Salles, 2006). Again, the value of such a benchmark is to help identify the combinations of interventions that come closest to reaching important goals, from saving lives to improving the quality of education.

In the next section, I discuss the basic steps and issues involved with estimating costs and calculating cost-effectiveness ratios. I focus especially on three "timing issues" that require particular attention: at what age effects on students begin to "count," the decay or compounding in effects over time, and discounting of costs and effects that arise in the future. While the main purpose in discussing these three issues is to facilitate the calculation of valid decision-making benchmarks, the approaches I propose also advance the methodology of cost-effectiveness research in general.

In the third section, I apply the above cost-effectiveness approach to previous empirical evidence related to class size, pre-kindergarten, computer-assisted instruction, cross-age tutoring, increasing instructional time, and whole school reform. In addition to providing the beginnings of an empirical basis for the benchmarks, discussion of these examples highlights the issues involved when comparing costs and effects across interventions and substitute categories. One key assumption required with the cost-effectiveness benchmarks is that policymakers face fixed budgets. Therefore, I also present

alternative benchmarks that view the size of the benchmarks from the perspective of their monetary benefits or societal "profitability," which is more appropriate in flexible budget contexts. The discussion also highlights some of the conceptual and methodological differences in the cost-effectiveness methodology that arise in education compared to research on health, where cost analysis is much more common.

Costs and Cost-Effectiveness Ratios

The discussion below provides a brief introduction to the concept and measurement of costs, the calculation of cost-effectiveness ratios and the three timing issues. For more in-depth discussions of many of the basic issues associated with cost concepts and measurement, the reader is referred to excellent texts such as Levin and McEwan (2001).

Conceptualizing and Measuring Costs

Economists define the costs of resources as the value of a resource in its next best use—the "opportunity cost." For example, for each hour a teacher spends instructing students, the teacher could have been working in some other job, spending time with her family, or some other personally valuable activities. Likewise, a teacher spending time in one instructional task (e.g., teaching math) takes time away from other instructional activities (e.g., teaching music). Using the hour to teach children a particular subject in a particular way therefore comes at cost in terms of these various lost opportunities.

The time of teachers and other personnel is indeed the largest resource in education. The value of this time is measured in terms of the compensation paid by educational systems to personnel because compensation is assumed to reflect the opportunity cost of personnel time.[3] Compensation includes both salary and employer contributions to health care and pensions because the latter are equivalent to salary from an economic standpoint. *Employee*

contributions to health care and pensions are not counted as compensation because they reflect consumption paid for by the employee rather than a market exchange between employees and employers. Some specific estimates of fringe benefits are provided later.

The fact that compensation can be used to measure the economic cost of personnel is intuitive from a practical decision-making perspective. When district administrators are considering a new program that would involve hiring more teachers, they will look at the budget and consider whether resources can and should be made available. In the case of teachers, the budgetary (or accounting) impact is often similar to the opportunity costs discussed above, but this is not always the case. Some budgetary costs over-state opportunity costs, e.g., if a state government decides to provide additional funding for school construction, it may pay for the schools over a 20-year period, but the value of the school, and thus its opportunity cost, is likely to last a half-century or more. Conversely, some economic costs may not show up in accounting costs. For example, if parental school choice is implemented and parents become responsible for transporting children to school, this would reduce the budgetary costs to the school district, but this is largely replaced by opportunity costs to parents. When there is a difference between budgetary and opportunity costs, it is recommended that researchers use opportunity costs because this includes all costs to society (Levin & McEwan, 2001; Monk & King, 1993).

Economists also distinguish between marginal costs, which vary by the number of students, and fixed costs, which are unrelated to the number of the number of students. Few costs are truly fixed in this sense. For example, a school building is not truly a fixed cost because the amount of space needed is clearly related to the number of students. However, school buildings and other costs are "lumpy" so that the cost per student still varies

somewhat with the number of students. Levin and McEwan (2001) recommend that, in these cases, multiple calculations should be provided for different levels of scale (i.e., different numbers of students).

While relatively low as a percentage of all educational expenditures, school buildings and other capital costs can be substantial in specific interventions. In these cases, costs can be annualized using a variety of methods (Levin & McEwan, 2001). Administrative costs (administrative assistants, accountants, etc.), which are often shared across programs being studied, are sometimes ignored if they can be considered fixed and independent of the program; alternatively, they can be apportioned based on the total instructional costs. Some educational interventions (e.g., school choice) may involve changes in administrative costs (e.g., student transportation) and, in these cases, administrative costs are obviously of central importance.

The time of teachers and parents represent important opportunity costs because these groups have alternative opportunities for work and leisure. For the same reason, the time of students is *not* generally counted as an opportunity cost. In addition to having few labor market opportunities, children in developed nations are legally required to be in school and expected to use their time for learning.[4]

Ideally, it would be possible to obtain a cost estimate for each individual student, based on observations of the actual number of units of each resource received (e.g., number of teacher hours), to estimate the opportunity cost of resources in perfectly competitive markets, and to calculate the cost by multiplying the resource units by their prices. In reality, researchers rely on prices from imperfect markets and, usually, have estimates only of the average resource units received by classrooms or schools rather than individual students. It

is also common, as the later examples highlight, to make back-of-the-envelope calculations using rough estimates of resources and/or using prices taken from broader samples (e.g., the average teacher salary in the state or nation). Not surprisingly, then, cost measures are subject to the same general types of measurement and sampling issues at outcome measures. Opportunity cost is the construct of interest and there are likely to be errors both in the resources units and their prices. There is also sampling error, though the standard errors are rarely reported.[5]

Calculating Cost-Effectiveness Ratios (CERs)

Cost-effectiveness analysis involves combining information about costs with information about effects. Cohen's *d* is the most common approach to measuring effect sizes:

$$d = (\mu_1 - \mu_2) / (\sqrt{(\sigma_1^2 + \sigma_2^2)} \quad (1)$$

where μ_i refers to mean values of the educational outcomes of interest for the treatment and comparison groups (indexed by *i*) and σ_i refers to the standard deviation of the outcome measure for the corresponding sample. In cases where effects are being estimated with coefficients from multivariate analysis, researchers often simply divide the coefficients in means by the overall sample standard deviation, which is equivalent to (1).

While the purpose of effect sizes is to facilitate comparisons of effects across interventions using different measurement instruments, this approach is far from perfect and some researchers have expressed considerable caution about their meaning and interpretation (e.g., Hill, Bloom, Black & Lipsey, 2007). For example, the standard deviation of student test scores tends to be higher for students in higher grade levels (e.g., high school). This means that, even if actual effects were identical between two interventions at different grade levels, the effects would appear larger in the lower grades because of the smaller standard

deviation. Similar problems arise when using other test scales such as grade level equivalents. Also, because the standard deviation is usually taken from the participating sample of students, the standard deviation and the effect sizes depend on the diversity of achievement in the sample. An intervention targeted to a specific group of students is likely have a smaller standard deviation than a broad-based sample, making the effect sizes in the targeted interventions appear larger. Because this study is focused on combining costs and effects, I refer the reader to other excellent discussions about effect sizes (e.g., Hill et al., 2007; Valentine and Cooper, 2003). The key point here is that, while these limitations are very important and much more research is needed to address them, making useful generalizations from educational research for policy purposes requires comparisons across studies and these comparisons require establishing a common unit of measure—the effect size is still the best, albeit flawed, measure reserchers have.

As described earlier, the cost-effectiveness of any intervention can be measured by the relationship between costs and effects. I focus, specifically, on the cost-effectiveness ratio, defined here as *the Cohen effect size divided by the total cost* (summed over all years), as shown in equation (2).

$$CER = d / C \tag{2}$$

As with effect sizes, larger cost-effectiveness (*CER*) ratios reflect more positively on the intervention. Because both the effect size and the cost estimates are subject to measurement and sampling error, the ratio of the two measures could have considerable error.

A variety of common issues arise when trying to calculate individual *CER*s and compare them across interventions. First, it is important to measure costs and effects in a common unit, nearly always on a *per student* basis. From the researcher's standpoint, it is

perhaps easiest to think of the costs of an experiment, so that the costs and effects relate to the differences in resources and outcomes received by the treatment group of students compared with the control group. Also, note that the effect size and costs should pertain to the same intervention, as applied to the same sample of students. For example, if an effect size is based on an achievement test given to 4th grade students at Smith Elementary who received a new type of reading instruction, then the costs should be measured for that specific reading instruction intervention implemented in that school with those students. As indicated above, some less-than-ideal methods to measuring costs involve using data about resources units and prices from separate or broader samples of students, classrooms, and schools.

It is important to adjust for inflation because of the changing value of money over time. All costs in this study are expressed in 2007 dollars, unless otherwise noted.[6] Thus, ideally, costs should be expressed in *real (inflation-adjusted) dollars per student*. In a similar vein, economists of education have identified variation in the costs of education across geographic regions, e.g., across states or between urban and rural areas. Taylor and Fowler (2006) report school district-level level indices ranging from 0.7-1.2 nationally. This means, for example, that a program in the lowest cost district would cost 70 percent of the value in the average cost district. While adjusting for such differences is helpful, there is no ready database for identifying geographic costs in specific locations. For this reason, and because the adjustments will typically be small, it may be impractical to recommend this as part of standard cost-effectiveness practice at the present time. However, making the geographic cost adjustments is very important in the benchmarking process (as discussed below).

Timing Issues: Age Counting, Decay/Compounding, and Discounting

At what age does an effect "count"? Education is commonly conceived as an investment in the future. From a research standpoint, this means that we are usually concerned with the long-term effects that education has on students, but researchers typically do not grapple with what counts as long term. For many outcomes, age 18 is a logical choice because this is typically the beginning of adulthood as people begin to vote, enter the workforce, and consider having children. This is therefore the point at which most of the effects of education, especially those related to academic outcomes, become valuable to society. For other student outcomes, especially juvenile delinquency and teenage pregnancy, the assumption that the effects begin to count at age 18 is undesirable.[7] Because counting short-term effects is much easier than counting long-term effects, I focus the later on age 18 as the age at which effects count.

Decay/Compounding. The fact that we are primarily interested in long-term effects, but typically only measure short-term ones, is problematic because there is no guarantee that the two will be the same. On the one hand, effects may decay. Students might forget what they learn and/or some interventions may be more effective than others in facilitating deep understanding that is remembered over the long-term. The evidence related to class size reduction and other interventions discussed later provide evidence that decay does occur.[8] In health research, the analog is that people who are initially cured of a disease such as cancer may experience a relapse in the future so that the initial effect of an intervention is only short-term in the sense discussed above. On the other hand, effects could compound; that is, skill begets skill. Improving reading skills in the short-term might make it easier to learn subsequent reading skills so that the long-term effect compounds and therefore grows over

time. Cuhna and Heckman (2007), for example, find evidence of compounding with respect to the impact of education on labor market earnings. It is possible that some outcomes are typically subject to decay and others are typically subject to compounding.

Regardless of whether effects decay or compound, it is important to distinguish *CER*s based on how far in the future the effects are measured. I define "short-term effects" as those measured near the end or immediately after the intervention ended, "medium-term effects" as those measured 1-4 years after the intervention, and "long-term" effects as those measured five or more years afterwards. The long-term effects are preferable because they are most likely to be permanent and not subject to further decay/compounding. Because of the possibility of decay and compounding, *CER*s should be compared only with *CER*s of the same term length.

While the main purpose of this section is to lay the groundwork for the benchmarks, it is worth noting that the issues of decay/compounding and "age counting" are rarely raised in the small number of cost-effectiveness analyses that do exist or in standard textbooks on the topic. The approaches proposed in this section therefore represent additional contributions of the present study.

Discounting. The third timing issue, discounting, is arguably the most familiar and commonly discussed in textbooks. However, the issue of "age counting" also requires some alteration in the typical discounting formulas.

There is broad agreement among economists and practitioners of cost-effectiveness analysis that costs and effects occurring in the future should be "discounted," or given less value, than those occurring in the present (Levin & McEwan, 2001; Lipscomb, Weinstein & Torrance, 1996). Intuitively, if a person has a dollar today, then she can put it in the bank

and it will grow at the rate of interest, whereas if she receives the dollar tomorrow, the potential to receive interest is diminished—the dollar tomorrow is therefore less valuable than one today (Levin & McEwan, 2001). While the effects are measured in monetary terms (dollars) in this example, the same logic also applies to non-monetary effects such as student achievement. The fact that some effects cannot be measured in monetary terms should not affect the weight that we give to outcomes at different points in time. We must therefore discount the value of all costs and effects that occur in the future, using the same rate of discount in each case (Lipscomb, Weinstein & Torrance, 1996; Muennig, 2002). Because effects occur at least as far in the future as costs in education, discounting reduces the measured value of effects at least as much as costs.

The earlier discussion argued that effects of education do not typically count until age 18. This means, for example, that if a one-year intervention occurs at age 12, then the effects of that intervention do not count for another six years. This is reflected in the following general formula for the discounted *CER* (CER^D) which is expressed as follows:

$$CER^D = \frac{d}{\overline{C}} \cdot \frac{\sum_{t=18-a}^{65-a} (1-\delta)^t}{\sum_{t=1}^{T} (1-\delta)^t} \tag{3}$$

where $\overline{C}$ is the annual cost of the intervention per student, δ is the discount rate, T is the length of the intervention (in years), and a is the age of the students when the intervention started. The expression of costs in *annual* terms ($\overline{C}$) in equation (3) is important to note because it is different from the *total* cost (C) in equation (2). Age 65 is chosen as the age at which effects stop counting. The choice of this age is somewhat debatable, especially as improvements in health allow people to live longer and with a greater quality of life.

Nevertheless, the choice of the ending age is much less important than the choice of starting age because, by age 65, the effects have already been discounted so much that alternative ages such as 70 or 75 have little impact on the CER^Ds.

I refer to the right-hand term of equation (3) as the "discount factor." Table 1a provides the discounting factors associated with different types of K-12 interventions based on equation (3), assuming a discount rate of three percent per year, which is the value recommended by a large number of studies including Barnett (1996), Lipscomb, Weinstein, & Torrance (1996), Moore et al. (2004), and Muennig (2002).[9] The horizontal axis indicates the number of years the intervention lasts and the vertical axis indicates the age and grade at which the intervention begins. Again, the discount factors are designed so that the undiscounted *CER* shown in equation (2) can be translated to the discounted *CER* simply by multiplying the undiscounted *CER* by the corresponding discount factor in Table 1a.

[Table 1a]

The smallest discount factor in Table 1a is for the intervention starting at age zero (infants) and lasting only one year. There are two reasons for this: (1) the numerator of the discounted *CER* is small because the effect occurs far in the future; and (2) the denominator is large because costs of such a short intervention are discounted little.

While there is considerable agreement about how discounting should be carried out in policy analysis, and that three percent is an appropriate baseline rate, researchers typically recognize that there is some uncertainty in the appropriate value. Quantifying the degree to which society values the present over the future is, not surprisingly, a difficult task and different approaches yield somewhat different results. Laxminarayan, Chow and Shahid–Salles (2006) therefore recommend considering a range, from zero to seven percent, as a

form of sensitivity analysis. Table 1b below provides the discount factors using seven percent. When the discount rate is zero percent (the future is just as important as the present), the discount factor is simply the number of years for which the effect counts. Starting at age 18 and ending at age 65 means that the discount factor is 47.[10]

[Table 1b]

Note that discounting is necessary because people prefer consumption in the present over consumption in the future, whereas inflation adjustments are necessary because the value of money changes over time. The inflation adjustments should be carried out before the application of the discount factors shown below because inflation rates vary by year, whereas discounting involves applying the same rate in each year.

Alternative discount factors are necessary for cases in which effects before age 18 are considered socially valuable (see earlier discussion of "age counting"). Also, for interventions that go beyond high school (e.g., adult education and government job retraining programs), the starting point for counting the effects would no longer be the same across interventions and the *CER*s would be affected in different ways by discounting. I have therefore created alternative tables of discount factors for these situations (available upon request).[11] These discount factors for post-high school interventions are smaller because the effects count for fewer years.

One concern that may arise from this discussion is that discounting appears to give an "unfair" advantage to interventions that last for many years or are aimed at older students. In some sense, this is true. If the most cost-effective interventions for older students, absent discounting, are exactly as cost-effective as the most cost-effective interventions for younger students, then the interventions for older students will be preferred because the value to

society from older students arises more quickly. Likewise, the costs of longer interventions are effectively lower because some of the costs are pushed further into the future and discounted accordingly.

This section has pointed out both the basic steps as well as more complex that arise in calculating cost-effectiveness ratios. Because an objective of this study is to make cost-effectiveness analysis more widespread, I have also boiled the complexities down to make the process manageable. The main task of the researcher is to estimate a valid effect size, which is still arguably the most difficult part. Converting this into a *CER* requires only adding up the costs, adjusting for inflation, calculating the undiscounted *CER*, and multiplying by the appropriate discount factor in the tables provided.

Cost-Effectiveness of Specific Interventions

The concepts and methods discussed above are applied in this section to a series of detailed examples. The selected examples are frequent subjects of debate in the policy arena and valid evidence about their costs and effects is available. Random control trials, or experiments, are widely considered to be the gold standard in scientific research because these studies avoid the possible selection biases that arise in correlational and some quasi-experimental studies (Shadish, Cook & Campbell, 2002). The review in this section therefore focuses on experiments, though a small number of longitudinal regression-based cost-effectiveness studies have been conducted (e.g., Grissmer et al., 2000; Harris, 2002).

This review and analysis is meant to be illustrative rather than comprehensive. It focuses only on the student outcome of academic achievement because this outcome is important to policymakers and, not coincidentally, because it is the only outcome for which there is remotely sufficient evidence to even begin the process of identifying benchmarks.

Studies of some of the interventions discussed below include measures of other important outcomes such as graduation from high school and non-cognitive outcomes.

In order to illustrate how cost-effectiveness benchmarks might be developed, I average the effects across academic subjects. This averaging is mainly to keep the amount of evidence to a manageable level for the development of the hypothetical benchmarks that comes later. Averaging across academic subjects is also warranted when evidence on the same set of subjects is generally available for multiple interventions so that the averages still represent measures of the same composite construct. The importance of using a common construct and metric cannot be over-emphasized.

Though most of the studies considered are based on experiments, this does not mean that the true effect sizes are agreed upon by researchers. In cases where concerns are raised about the validity of the effects, I therefore create upper- and lower-bound estimates as a form of sensitivity analysis. This section does not carry out any new cost-effectiveness analysis, except to convert the results from previous studies into common and comparable *CER* ratios. All costs are converted to 2007 dollars.

Evidence on Class Size

The Tennessee STAR class size experiment took place in the 1980's and included random assignment of 12,000 students to small and large classes for one or more years in grades kindergarten through three (K-3). The average large class had approximately 22 students and the average small class had 15 students. The program was not targeted to a specific student population and participating students came from a wide variety of academic and family backgrounds. Concerns about the validity of the STAR results are discussed by Grissmer (1999), Hanushek (1999), Harris (2000, 2002) and Krueger (2003).

Nye, Hedges, and Konstantopolous (NHK, 1999) report a variety of effect size estimates at multiple grades (3, 4, 6, and 8) and for multiple subjects (math, reading, and science). The effects below are also averaged across subjects (within grades) and it is these averaged effects that are used as the basis for the calculation of *CER*s. The results for grades 3 and 8 are considered short-term and long-term, respectively. The results for grades 4 and 6 are averaged together to obtain the medium-term results.

Throughout this discussion, d_{SU} refers to the short-term (subscript *S*) upper-bound (subscript *U*) effect size estimate, d_{MU} refers to the medium-term upper-bound estimate, and so on for d_{LU}, d_{SL}, d_{ML}, an d_{LL}. Based on a comparison of students who were in small classes for all four years versus STAR control classrooms, this yields d_{SU}=0.379, d_{MU}=0.365, and d_{LU}=0.331. Note that the pattern of declining effect sizes here, and with other interventions below, provides some evidence that decay occurs.

The effect sizes drop off considerably when the estimate is based on the initial assignment to control and treatment groups (intent-to-treat, ITT). On the one hand, these effects are likely biased downward if there is a class size effect because students who were not treated, but still counted as treated, will pull down the treatment post-test. On the other hand, the treatment-on-treated effects may over-state the true effects if there is differential attrition and therefore represent an upper-bound. Again, it is beyond the scope of this study to resolve these issues. I use the ITT estimate as the lower bound for the STAR short-term and medium-term effects (d_{SL}=0.160, d_{ML}=0.143). All of these estimates come from Table 6 in NHK (1999).

Krueger and Whitmore (2001) report even longer-term effects on college entrance exam scores. They find that the effects may diminish further in later years, but that there are

still statistically significant differences of 0.120-0.130 standard deviations nine years after the policy was completed. This is similar to the long-term estimate from the ITT estimates (0.166).[12] I therefore choose d_{LL}=0.120.

Throughout this section, the range of reasonable effect size estimates is based on the lower- and upper-bounds above. The short-term range d_S is formed from the lower- and upper-bounds (d_{SL} and d_{SU}) and so on for d_M and d_L. Using this notation, the ranges of effect size estimates for STAR are: d_S=0.160-0.379, d_M=0.143-0.365, and d_L=0.120-0.331. Borman and Hewes (2002) reported "sustained effect sizes" of d=0.140 (reading) and d=0.190 (math) for an average of 0.175, which is near the lower end of the long-term range above. Estimates of the effects of other class size interventions are discussed by Grissmer (1999) and Harris (2000, 2002).[13] Because these estimates from studies of other class size reduction initiatives are within the above range for STAR, they are not discussed further.

Costs of small classes. The main cost of class size reduction is additional teacher time. According to the U.S. Department of Education (2006), the projected average teacher salary in 2007 is $47,017 in 2004 dollars, yielding $51,444 in 2007 dollars, excluding fringe benefits. According to the U.S. Census (2005), fringe benefits for school instructional staff are 23.9-27.3 percent above the salary level. The exact number depends on the composition of "other compensation" that the document does not report. Podgursky (2005) reports that fringe benefits are 20.2 percent of total compensation requiring a 25.3 percent upward adjustment in salaries, which is within the range from the Census data (23.9-27.3). I use Podgursky's figure, yielding total average annual teacher compensation of $64,459. This value of teacher compensation is used throughout the remainder of the analysis.[14]

In a school with 100 students, a class size of 22 would require 4.54 teachers, compared with 6.67 teachers for a class size of 15. This implies that the reduction requires 2.13 additional teachers. (The percentage change of 47 percent would be the same no matter the total number of students.) The main cost of STAR is therefore the cost per teacher \$64,459 times 2.13 teachers, yielding costs of \$137,298 or \$1,373 per student per year. Because STAR involved four consecutive years of class size reduction, this is multiplied by four to obtain \$5,482 (undiscounted). This excludes the cost of classroom space, I therefore add to this a capital cost estimate of \$764 per student based on Harris (2004),[15] yielding \$6,256 per student for four years of participation in small classes in STAR.

Throughout this section, I define C_L as the lower-bound cost estimate and C_U as the upper-bound cost estimate. Because costs do not "decay" in the same sense that effects do, there are no "short-term" or "long-term" costs and therefore no need for the accompanying subscripts. Based on the above figures for class size, C_U=\$6,256. These costs are considerably larger than the costs of STAR reported by Borman and Hewes (2002) who use a figure of \$3,641. The costs from Borman and Hewes exclude capital costs; therefore, I add the estimated capital costs above (\$764) to the \$3,641 to obtain C_L=\$4,405. Brewer, Krop, Gill, and Reichardt (1999) also estimate the costs of class size reduction, but they focus on how the costs vary in different types of class size interventions. Because their estimates cannot be directly connected to any particular effect size, the Brewer et al. study is not discussed further.

Based on the above ranges of estimates for effects and costs, I identify a reasonable range of *CER*s. The lower-bound short-term *CER* is obtained by dividing d_{SL} by C_U, while upper-bound short-term *CER* is d_{SU} divided by C_L and so on for the medium-term and long-term

estimates. For STAR class size, this yields: CER_S=0.026-0.086, CER_M=0.023-0.083, and CER_L=0.019-0.075 (see Table 2). The above results are similar to the *CER*s for STAR reported by Harris (2002) as 0.031-0.048. This implies that for a cost of $1,000 per student (undiscounted), policymakers can obtain short-term effects of 0.026-0.086 standard deviations, and so on for medium- and long-term effects. The undiscounted *CER*s are multiplied by the discount factors in Table 1a (δ=0.03) to obtain the discounted *CER*s, shown in the right-hand columns of Table 2. It is important to emphasize that there are a wide variety of interventions listed in Table 2 and that, for reasons explained below in further detail, only some of them can be compared with one another.

[Table 2]

Evidence on Early Childhood Education

There has been widespread interest in recent years in programs for children who are less than the typical school age. This is partly based on evidence from two early education programs that have shown evidence of long-term effects and cost-effectiveness. Early childhood programs can vary considerably in the starting age (usually age 3 or 4), the number of years children participate (usually one or two years), the amount of time participating children spend in the program (e.g., half-day versus full day), and the quality and nature of the care/instruction provided (e.g., small versus large classes; extensively trained versus modestly trained teachers). These variations result in differences in costs, as well as perhaps effects.

Abecedarian

As described by Campbell et al. (2001), the well known Abecedarian Project involved 111 infants in one of four cohorts in the years 1972-1977 who were randomly assigned to

control and treatment groups at an average age of 4.4 months. All of the infants were born to low-income families and 98 percent were African-American. The treatment involved high quality, full-time child care for 50 weeks per year. The average child-teacher ratio was 3:1 for children up to age four and the ratio averaged 6:1 for children of age five. The program had a formal curriculum, which focused on language development and pre-literacy skills as students progressed through the program.[16]

The effect size for achievement at age 15 is reported by Campbell et al. (2001) as d_L=0.410 (averaged across math and reading; there are no lower- and upper-bound estimates). Short-term and medium-term effects on achievement are not reported. However, the initial effect size on "cognitive skills" are reported as d_S=0.740. Even though cognitive skills are conceptually different from achievement, the decay seen in Tennessee STAR class size effects suggests that this figure may be a reasonable approximation of the short-term achievement effect, though the uncertainty over whether effects typically decay or compound makes this far from definitive.

Costs of Abecedarian. The costs in the case of early childhood education are different in one important way from the calculation of costs in class size and most other policies because early childhood education substitutes one set of inputs—teachers, etc.—for another input—parental time or other child care. This is not an issue in studying K-12 because students in most of these grades are legally required to be in school and the choices center on what resources should be provided while they are in school. Early childhood education, in contrast, is optional so that the measured costs of parent time saved needs to be subtracted from the school resources in measuring the total costs. This is a good example

where the accounting costs of an intervention can differ substantially from the opportunity costs.

The direct costs of Abecedarian are estimated by Masse and Barnett (2002) as C_L=$43,955. They indicate that they use the method recommended by Levin and McEwan (2001). While they do not show the costs or how they were aggregated, Masse and Barnett appear to make the appropriate adjustments, including the cost savings to parents. These costs are somewhat lower than the estimates used in Borman and Hewes (2002) who cite a figure of C_U=$62,990. As shown in Table 2, the above estimates of effects and costs yield CER_S=0.012-0.017 and CER_L=0.007-0.009.

Perry Preschool

Even more widely cited than the Abecedarian Project is the Perry Preschool Project. An experimental design was used and participants were followed over a number of years. The treatment group received half-day pre-kindergarten, five days per week during eight-month periods (October through May) during the years 1962-65. Most of the students entered at age 3 and participated for two years, though the initial cohort entered at age 4 and participated for only one year. The pupil-teacher ratio averaged approximately 3:1 across cohorts and years. In addition, teachers made a 1.5-hour home visit to each child once per week. The intervention was targeted to disadvantaged students. All students were African-American and had low IQs and mothers with very low levels of education (Barnett, 1992). All teachers were experienced and certified in both special education and preschool education.

Barnett (1992) provides evidence on the effects of the project on achievement and intelligence. While no long-term effects are found regarding intelligence, there are

statistically significant gains in achievement. At age 14, students in the pre-kindergarten treatment groups had achievement effects of d_L=0.480 standard deviations, which is similar to the "sustained effect sizes" of d_U=0.500 reported by Borman and Hewes (2002).[17]

Costs of Perry. Barnett (1985, 1992) provides data regarding costs. Again, the underlying ingredient costs are not provided, though it appears that parental time was taken into account. He reports total program costs of C_U=$21,961. Borman and Hewes (2002) calculate costs as C_L= $18,647. This yields CER_L range of 0.022-0.027 (see Table 2).

Computer-Aided Instruction, Cross-Age Tutoring and Instructional Time

Levin, Glass and Meister (1987) review evidence on the cost-effectiveness of computer-aided instruction (CAI), cross-age tutoring, and additional instructional time.[18] Their approach is to review other studies on the effects of policies and, because studies on these topics do not address costs directly, they estimate the costs of the interventions themselves from the available information (after the fact). The analysis of each intervention is based on a single study regarding effects. Only the study of CAI is based on any form of randomization, while the others are quasi-experimental methods. No information is available regarding the timing of the effect measures and it is assumed, as is typically the case, that the effects were measured immediately after the intervention (short-term effects).

Levin et al. calculate *CER*s for the various interventions under study, though the underlying cost and effects data are not provided. Converting their *CER*s to the format described above yields: CAI CER_S=0.600, Instructional time CER_S=0.400, Peer and adult tutor CER_S=0.900, Peer tutor CER_S=1.400, Adult tutor CER_S=0.300. As we will see, these short-term *CER*s are large compared with the other interventions.

Success for All

While the education interventions below change one element of instruction, whole school reform takes a different approach changing a variety of interconnected instructional and administrative elements. Several studies report *Success for All* (SFA) as being among the few comprehensive school reform efforts to have positive and statistically significant effects on student achievement (Borman and Hewes, 2002; Borman, Hewes, Overman, and Brown, 2003).

Borman and Hewes (2002) report d_L=0.200, based on a matched sample methodology. While this effect is based on an average across academic subjects, the larger effects they find for reading are sensible given the program's focus on literacy. The authors report that these are ITT estimates.

Borman, Slavin, Cheung, Chamberlain, Madden, & Chambers (2006) use an experimental design and report SFA effect sizes of d_{SL}=0.210 and d_{SU}=0.330. The lower value is similar to that from Borman and Hewes (2002). For this reason, and because the Borman et al. (2006) used a randomized design, the Borman et al. (2006) range is used throughout. (Note that one of the above evaluators, Slavin, is the founder of SFA and remains actively involved in the foundation that promotes and operates the program.)

Costs of SFA. Borman and Hewes (2002) report costs of $3,666 per student for SFA. The costs used by Borman and Hewes are smaller than those made previously by King (1994). She identifies two categories of costs: budgetary (accounting) and additional time of existing personnel and parents (in hours). To calculate the total costs, I calculated the monetary value of these time costs and added these to the budgetary costs. In contrast to the earlier calculation in which I largely rely on the cost estimates from the cited authors, I

describe the conversion of the King estimates in some detail as an example of how cost estimates can be made in practice even when, as in Levin et al. (1987), ideal observation-level cost data are unavailable.

King reports the additional time of existing personnel as 0.5-1.5 hours per week for all personnel and 4-11.5 hours per week for "some staff/parents," though the number of staff and parents to which these hours apply is unclear. "Some staff/parents" implies that the costs apply to less than half of the staff and parents. Below, I assume that the hours for "some staff/parents" apply to 25 percent of staff and parents. To estimate the value of parent time, I looked to the Bureau of Labor Statistics (2007) which reports the national median usual weekly earnings in 2006 for females aged 20-24 was $413 and for females aged 25-34 it was $583. This leads to a weighted average of $526 per week. Assuming 40 hours per week, this yields an hourly age of $13.16. Estimating fringe benefits for this specific group of students is difficult. However, it is surely less that the numbers mentioned earlier for teachers. I assume that fringe benefits require a 10 percent increase in earnings so that total compensation is $14.47 per hour. This provides the estimate of the cost of parental time if only mothers put in extra time in SFA.

The cost of staff time is calculated using the national annual teacher salary mentioned above ($64,459). Podgursky (2005) reports 186 contract days per year, implying 1,488 hours per year (assuming eight hours per day). This is surely an under-statement as teachers typically perform work activities at home during the week, on weekends, and during the summer, outside of the contract periods. I therefore assume 1,800 estimated hours per year (still considerably less than the average full-time worker), which yields $35.81 (2006 dollars), or $36.71 (2007 dollars). King's estimates are based on a school with 500 students.

With a pupil-teacher ratio of 16.2, this implies the school has 31 teachers. Rice indicates that the average SFA teacher adds one hour per week compared with other teachers. One additional hour for each of 31 teachers for 40 weeks per year, at $36.71 per hour yields $45,520 per school.

As indicated above, it is assumed that 25 percent of teachers also put in longer hours, beyond the one hour per week. In a school with 31 teachers, 25 percent of this figure yields eight teachers, each of whom requires eight additional hours per week for each of 40 weeks per year, yielding total annual teacher time worth $93,978. The parental time costs might seem small at first because it is only one hour per week, but there are many more parents than teachers. From King's estimates, assuming that 25 percent of families devote one additional hour of parental time per week for each of 37 weeks at $14.47 per hour, this implies a parental time cost of $66,924. (I use 37 instead of 40 weeks because parents are less likely than teachers to put in time during the summer.) This yields total non-budgetary costs of $45,520+$93,978+$66,924=$206,422, or $413 per student per year. King estimates the budgetary costs as $769-1,905 per student per year. Adding this to the time cost yields total costs per student of $1,182-$2,318 per student per year.

The Borman and Hewes estimate, at $954 per student per year, is below this range. One possible explanation is that King considers a high-fidelity implementation of the intervention whereas Borman and Hewes consider costs of actual implementation in which certain prescribed resources might not be utilized. Because the effects identified by Borman and Hewes are based on these actual resources, both sets of cost figures are arguably correct, depending on which effect the costs are associated with. I choose the Borman and Hewes estimates as the lower bound (C_L=$954) and my adapted estimates of the King data as the

upper bound (C_U=$2,318). Unlike the costs reported above, these are reported in annual terms because the *CER*s are calculated based on the different numbers of years involved in the different effect estimates, 3.84 years in Borman and Hewes (2002) and 3.00 years for Borman et al. (2005). This implies CER_S=0.030-0.116 and CER_L=0.022-0.055 (undiscounted). Again, all of the above undiscounted *CER*s are translated to their discounted forms in the right-hand columns of Table 2.

Discussion and Summary

These applications of cost-effectiveness analysis to class size reduction, early childhood education, peer tutoring, CAI, and whole school reform are some of the most well known and respected examples of cost-effectiveness and represent the types of research that cost-based benchmarks would hopefully encourage in the future. These examples also serve to illustrate the variety of issues that arise in calculating *CER*s: dealing with uncertainty about the actual effect sizes, estimating costs without student-level data, and distinguishing between budgetary and opportunity costs.

This empirical evidence also provides a basis for showing how to create cost-effectiveness benchmarks. Before taking this step, however, it is necessary to outline one final critical element of the cost benchmarks process.

The Decision-Making Context

To establish benchmarks that are useful for decision-making purposes, it is necessary to start with a clear decision *framework* and a clear outline of the decision *context*. The second section focused on the decision framework, which is rooted in the economics concepts of efficiency and opportunity cost, using cost-effectiveness ratios as the primary analytic tool. The discussion below turns to the assumptions about the decision context and

evidence about the assumptions. The objective here is to provide a sense of what is necessary for cost-effectiveness benchmarks to improve policy decisions and education outcomes.

Assumption (1). The decision-maker's objective is to maximize the expected level of student achievement. This is really just another way of saying that the decision-maker accepts the economics-based decision-making framework outlined above, with the additional practical requirement that the objective is to maximize student achievement. Student achievement tests have become increasingly common in recent years and increasingly accepted as a central measure of educational objectives, especially with the adoption of the federal *No Child Left Behind* policy requiring annual testing in grades 3-8. There is also evidence that student achievement is indeed the primary, tough by no means the only, objective of educators (Rothstein and Jacobsen, 2006).[19] Another advantage of using student achievement is that it can be measured regularly while students are still in school, facilitating the measurement of effect sizes.

A disadvantage of focusing on achievement scores is that they may be only weakly related to larger objectives of education such as social cohesion, civic participation and worker productivity. Ideally, this could be addressed by creating a weighted index of the multiple outcomes. In reality, different studies measure different outcomes (partly because they have different objectives). This creates a difficult problem. Perhaps the only practical solution is to create separate benchmarks for each outcome and then leave it to the discretion of the decision-maker to subjectively weigh the collection of evidence.[20]

Assumption (1) also requires that the decision-maker is maximizing the *expected value* of student achievement. This implies that the decision-maker is risk-neutral so that two

policies with identical effect size point estimates are equally valued. For example, if the cost-effectiveness ratios for interventions A and B are *CER*=0.05, but the ratio for A is certain and the ratio for B has an equal probability of *CER*=0 and *CER*=0.10, then the expected CER in both cases is 0.05 and the decision-maker has no preference between A and B. To the degree that this assumption is valid, it means that the standard error of the *CER* is not a central issue to the decision because the point estimate represents the expected value.

First Corollary to Assumption (1): The equity of educational outcomes is unimportant. The benchmarks here assume that the objective is to maximize average student achievement, which implies that equity is unimportant. However, this framework is easily adapted to decision contexts where equity is important. We could use this same evidence to consider the question, what is the most cost-effective way to raise achievement for disadvantaged students? Framing the question in this way is arguably superior to other recent attempts to measure the size of test score effects for minorities in terms of the "achievement gap" between minority groups. For example, Peterson and Howell (2006) argue that the achievement effects they find from school vouchers are large enough to substantially reduce the racial achievement gap. Again, even if an intervention reduces the achievement gap substantially (and there are reasons to question this in the case of vouchers[21]), decision-makers must still consider the costs of doing so and will be more likely to reduce the achievement gap if they look for the most cost-effective approaches to reaching that objective. The cost-effectiveness criterion is a flexible one that can be used to identify appropriate interventions for a wide range of educational objectives, including equity.[22]

Second Corollary to Assumption (1). The distribution of costs among stakeholders is unimportant. Levin and McEwan (2001) in their textbook on cost-effectiveness analysis

discuss the importance of breaking down costs according to who is responsible for them, e.g., local versus state governments. Assumption (1) means that the distribution of cost, like the distribution of effects mentioned above, is unimportant to the decision-maker.

While all decision-makers are obviously quite concerned about whether they, or some other decision-maker, must bear the cost burden, it is also important to realize that political systems allow decision-makers to negotiate, so that a stakeholder bearing a substantial cost can negotiate for additional resources from other decision-makers who may bear few costs but who have much to gain by improving student outcomes. In this respect, the distribution of costs is less important than it first appears and, in any event, it is impossible to generalize about who bears costs across contexts and policy options.

Assumption (2). The generalizability problem is the same across interventions. External validity is one of the central criteria for evaluating the effects of interventions (Shadish, Cook, & Campbell, 2002), yet it is common that the effects found even in the most rigorous studies fail to arise in other contexts. One specific problem is that interventions get implemented differently at scale—i.e., as general policies versus small experiments. For example, as Harris (2007) points out, the Tennessee STAR experiment involved hiring more teachers, presumably from outside the study sample to fill the new positions. It is therefore possible that the students in the schools where the teachers came from learned less as a result of their teacher's departure, but this negative effect would not be captured in the study. Yet, when implemented at scale (e.g., the statewide programs in California and Florida), these negative effects are captured and likely reduce the true impact of policies. This issue also arises on the cost side of the equation: hiring teachers for the STAR experiment involves

relatively few teachers, but statewide programs can place demands on resources that are quite different—sometimes smaller, sometimes larger—than those measured in the study.

This problem may be smaller than it appears if all the effect sizes used to create the benchmarks are similarly affected by the scale-up problem. But this requires a leap of faith and there is really no way to know the nature of the scale-up problem or other issues of generalizability until there are multiple rigorous studies, some based on scaled-up policy interventions. It is worth noting, however, that this problem is equally great when using the Cohen benchmarks and ignoring costs. The generalizability problem is not unique to cost-effectiveness.

Assumption (3). Decision-makers have a basic understanding of their student populations, policies, and other characteristics of the learning environment and can target resources accordingly. This assumption requires that decision-makers only consider whether to apply an intervention to the types of students for whom evidence is available, that the decision-maker is responsible for at least some students of those types, and that the decision-maker can target resources to students accordingly. This last issue is important because there is often political pressure to provide all programs to all students, which means that the ultimate decisions made are less cost-effective and therefore result in lower average outcomes. A good example of this problem is Florida's statewide class size reduction program that covers grades K-12, even though there is essentially no evidence that class size reduction is effective above grade 3.

Assumption (3) also requires that the interventions from which the benchmarks are derived are within the decision-maker's purview and therefore represent reasonable "substitutes." At the one extreme are, for example, school principals who have relatively

little discretion over school resources in general and, when they do have such discretion, are likely to be choosing from among very similar options—e.g., alternative math curricula. In this case, the benchmarks need to be based on "near substitutes," or alternatives that are similar with respect to the educational objectives, grade levels, subjects, student population, scale of intervention, and mechanisms to which the resources will be applied—what I call the six main "dimensions of substitutability."

Alternatively, district, state and federal policymakers can make policies about a broad range of policy alternatives. For example, over the past several decades, governments have adopted policies that reduce class sizes, rather than raise teacher salaries (Harris, 2007). As Harris points out, class size has typically targeted lower grade levels and involved hiring more teachers. The issue of salaries has been raised at all grade levels and generally means hiring fewer teachers. Thus, class size and teacher salaries are not just different, but arguably opposite, strategies and yet they are both within the purview of policymakers at multiple levels of government. As a general rule then, we can say that the wider the purview of the decision-maker, the broader the range of interventions for which evidence should be brought to bear and the more important it is to consider far substitutes such as class size and teacher salaries. Later, I discuss some specific issues that arise in developing benchmarks for near versus far substitutes.

This is another case where research from health research is instructive. The World Bank's World Development Report (1993) is one of the most well known, comprehensive efforts to identify cost-effective health policies. This report, and more recent efforts such as the World Bank's collaboration in the Disease Control Priorities Project (DCPP, http://www.dcp2.org/main/Home.html), have used cost-effectiveness analysis to identify

both broad priorities (equivalent to "far substitutes") and different approaches to treating specific diseases such as cancer (equivalent to "near substitutes"). For example, one recent DCPP report argues that "interventions to treat communicable diseases have been highly cost-effective in the past and remain so despite new challenges, such as drug-resistant pathogens" (Laxminarayan, Chow, & Shahid–Salles, 2006, p.53). By implicitly contrasting improvements in health aimed at communicable with those aimed at non-communicable diseases, the authors illustrate why comparing far substitutes is so important. A doctor specializing in cancer treatment might be interested only in alternative cancer treatments, but broader policymaking bodies such as the World Bank have a clear need to know whether the general strategies are likely to be most cost-effective. This is equivalent to the difference in education between school principals who have relatively little control over school resources and state legislators who, like the World Bank, has broad purview over educational options within their jurisdictions.

A final piece of Assumption (3) is that the policies on which the "large" benchmarks are based have not already been implemented in the specific decision-making context. The logic of the benchmarks is that an intervention should be adopted if its *CER* ratio is larger than that of all the other options. But if a specific school district or state has already adopted the policies with the largest *CER* ratios, then it might be sensible to consider policies with *CER* ratios that only meet the "medium" benchmarks. This means that decision-makers must have basic knowledge of the policies they have implemented.

Assumption (4). The decision-maker faces a fixed budget. The assumption of a fixed budget is most relevant to school and district leaders who have little control over budgets. In this case, the best course of action is to allocate those resources to the most cost-effective

programs, as indicated above. However, it is reasonable to argue that this is the wrong framework for decisions made at the state and federal level where funding is increasingly controlled and where policymakers can reduce funds or raise additional funds. In this case, the appropriate approach is an economic cost-benefit or "monetary benefits" approach where the costs and effects are compared with all other potential uses of resources, including other governmental programs and private consumption of goods and services, which would be reduced if the government had to raise taxes to pay for additional spending.

The monetary benefits approach involves measuring educational benefits in monetary terms and subtracting opportunity costs to obtain a measure of societal "profit." The most commonly studied monetary benefits among economists are those that arise from higher earnings of workers in the labor market. In addition to being one of the easier education benefits to measure in monetary terms, this measure emphasizes the idea that workforce development is an important role for formal education.

Again, these four main assumptions focus on the nature of the decision-making context. In one case (maximizing achievement), the assumption is supported by evidence, while other assumptions (e.g., that budgets are fixed) can be relaxed. The assumption about the decision-maker's understanding of the local context is not easily tested and, in any event, this is required for making any type of generalization.

One can certainly debate to what degree these are realistic, but it is important to emphasize that their realism must be considered in relation to typical decisions. In other words, we must ask the question, will better decisions result from the benchmarks on which these assumptions are based or from more typical decision-making processes that exclude systematic consideration of costs? Given the centrality of costs to decision-making, I argue

that the presence of benchmarks and the consequent expansion of cost-effectiveness research is very likely to improve decision-making.

While this section has focused on the decision context, it is worth noting two additional assumptions about the benchmarks themselves that are interrelated with the contextual assumptions. First, the costs and effects of each intervention must be independent of the other interventions. Thus, for example, the effects of class size reduction are assumed to be unrelated to the amount of peer tutoring student receive. Second, the benchmarks must be sufficiently detailed—we might say "complete"—to allow decision-makers to take into account the unique characteristics of their contexts. While this is certainly not the case at present, the purpose of this study is to move research toward a complete set of benchmarks. We move one step further in that direction in the next section.

Toward Cost-Effectiveness Benchmarks

The next and final step in the benchmarking process is to use evidence to create the benchmarks. While the evidence considered here is necessarily too brief to create actual benchmarks, the evidence on class size, etc. can be used to illustrate the process and some of the problems that arise. I start by using the above evidence to create hypothetical benchmarks for far substitutes, followed by a discussion of near substitutes. Then, I relax the assumption of the fixed budget and create preliminary benchmarks based on the assumptions of flexible budgets, using a monetary benefits approach.

Hypothetical CER Benchmarks for "Far Substitutes"

As discussed in the second section, the basic economic decision framework requires that the most cost-effective policies be adopted by decision-makers before less cost-effective policies. This means that a large *CER* is one that is larger than all of the others shown in

Table 2. By that standard, a policy would have to have a short-term *CER* that is larger than 26.18 (the discounted *CER* for peer tutoring). An intervention with a *CER* ratio of this size is more cost-effective than any the others discussed and, therefore, a good candidate for adoption.

At least two potential problems could arise if this rule were applied literally: First, a single large outlier *CER* might be given undue influence over the benchmarks. The outlier problem looks potentially significant when considering Table 2 as there are significant drop-offs in the size of the *CER*s, from 26.18 (the highest of three peer tutoring options) to 11.57 (computer-aided instruction) to 7.716 (instructional time) and 0.544-2.104 (Success for All). If we were to define 26.18 as a large benchmark, none of the other interventions would be even close to reaching it.

Second, because many interventions have *CER* evidence only for one or two term lengths, the benchmarks at each term length are necessarily based on evidence from a different set of interventions. This is undesirable in itself, but could also result in illogical patterns in the benchmarks. The first row of Table 3 shows the highest *CER* in each term length. The largest discounted short-term *CER* (26.18) is roughly 20 times larger than the largest long-term *CER* (1.381). This occurs mainly because no evidence is available for medium-term and long-term for those interventions that are apparently most cost-effective in the short-term evidence (peer tutoring, etc.). Thus, the largest *CER*s in the medium- and long-terms come from interventions that were probably only moderately cost-effective in the short-term, though this interpretation is untestable.

One way to approach this problem is to extrapolate large *CER*s across term lengths. Three of the included studies have *CER*s at different term lengths: STAR, Abecedarian, and

SFA. While they all appear to display decay, the maximum decay rate among these three occurs with Abecedarian where the effect size drops from d_S=0.740 to d_L=0.410, a decline of 45 percent. Based on this maximum decay rate (rounded to 50 percent), a long-term peer tutoring *CER* of 1.381 can be estimated, still nearly ten times larger than the largest actual long-term *CER*. As a result of this and the other potential problems, translating the evidence from Table 2 to benchmarks is necessarily a combination of art and science.

[Table 3]

It is useful to start with the short-term discounted *CER*s because this is where the most evidence is available in general and in Table 2. Notwithstanding the precipitous drop-off mentioned above, 6.0 might be considered a reasonable estimate of the "large" short-term benchmark—only the various peer tutoring options and instructional time reach this level. There is another group of estimates in the 0.479-2.104 range. Therefore, 0.4-6.0 is a possible "medium" short-term *CER* and less than 0.4 could be considered "small" for a short-term *CER*. At least one of the *CER* ranges for specific interventions in Table 2 fit into one of the three hypothetical *CER* benchmark categories in Table 3. Similar and slightly smaller benchmarks are listed in Table 3 for the medium- and long-term hypothetical benchmarks, based on the extrapolated *CER*s discussed above. It is important to emphasize again that these interventions are "far substitutes," meaning that they involve comparisons of interventions that vary with regard to dimensions of substitutability.

Comments on Potential Benchmarks for "Near Substitutes"

While some policymakers might be interested in far substitutes, others have narrower purviews and are therefore interested only in "near substitutes." For these decision-makers, it is necessary to create separate benchmark tables for each combination of dimensions. A

comprehensive effort in this regard will require substantial additional research. Even if we were to limit the discussion to different subjects and grade levels/ages, this would still result in at least 60 different tables (e.g., reading in pre-kindergarten, math in grade 3, etc.).[23] Adding different types of students, levels of scale, and mechanisms will easily lead to hundreds of tables. While the availability of rigorous research is growing, cost analysis remains in extremely short supply. Benchmarks in these cases might have to be based on only one or two other interventions, as they are in the health literature.

Of the examples considered above, Tennessee STAR class size reduction and *Success for All* (SFA) provide the most plausible pair of near substitutes. Both are aimed at students in the lower elementary grades, involve substantial additional resources, and are focused on raising achievement. Even in these cases, however, there are important differences. STAR involved as much 2-6 times the costs per student as SFA (see earlier lower- and upper-bound cost estimates) and the latter is much more explicitly focused on literacy than is STAR. This highlights the complexities of finding near substitutes and the usefulness of considering both near and far substitutes.

Benchmarks for Flexible Budget Contexts

Assumption (4) requires that the decision-maker face a fixed budget. This is not always the case, which means that an alternative set of benchmarks instead might be based on estimates of the monetary benefits of education, rooted in an economic method called cost-benefit analysis. As with cost-effectiveness, an excellent general introduction to this method can be found in Levin and McEwan (2001).

In order to identify alternative benchmarks that can be compared with those in Table 4, it is necessary to formulate the monetary benefits of education in terms of the Cohen effect

size *d* and to convert this to monetary terms by taking into account worker wages (*w*) and the labor market return to student achievement (γ). This leads to the general benefit function $B_t(\gamma,w,d)$.[24] Costs continue to be measured as discussed earlier and as shown in equation (3). The costs and benefits are reflected in equation (4):

$$\sum_{t=1}^{T}(1-\delta)^t\,\overline{C} = \sum_{t=18-a}^{65-a}(1-\delta)^t\,B_t(\gamma,w,d) \tag{4}$$

When equation (4) is satisfied and costs equal benefits, society's net benefits are zero. For more detailed discussion of equation (4), see Harris (2007), Krueger (2003), and Krueger and Whitmore (2001) who apply a similar cost-benefit approach to class size reduction.

Because γ is typically expressed as a percentage of annual earnings, the benefits of educational can be calculated simply by multiplying the three parameters in parentheses together; that is, $B_t(\cdot)=\gamma\cdot w\cdot d$. Substituting this into (4) and re-arranging terms yields:

$$\frac{1}{\gamma\cdot w} = \frac{d}{\overline{C}}\cdot\frac{\sum_{t=18-a}^{65-a}(1-\delta)^t}{\sum_{t=1}^{T}(1-\delta)^t} \tag{5}$$

The right-hand side of equation (5) is identical to the right-hand wide of equation (3). Thus, we can say that the discounted *CER*s also meet the flexible budget requirement if they are larger than $(\gamma\cdot w)^{-1}$. This naturally implies that the more positive the relationship between achievement and wages, the smaller is the discounted *CER* necessary in order to reach positive social profitability. (Because the discounted *CER*s in Tables 2 and 3 express costs in

thousands, a direct comparison with the earlier tables also requires multiplying $(\gamma \cdot w)^{-1}$ by $1,000 to achieve a common scale.)

The influence of achievement on students' future labor earnings is therefore the critical factor in determining the break-even *CER*s. Krueger and Whitmore (2001) assume γ=0.08, so that a one standard deviation in test scores is associated with an eight percent increase in annual earnings. In contrast, Krueger (2003) assumes that a one standard deviation increase in student test scores is associated with a 20 percent increase in students' future wages for each subject, yielding a "total" effect yields γ=0.40, which Krueger and Whitmore acknowledge is probably an overestimate. Evidence from the Perry Preschool Project, which considered effect on both achievement and earnings for the same sample of students for example suggests that γ=0.052. The value γ=0.08, used by Kruger and Whitmore (2001) is a middle ground and is the base value used to calculate the benchmarks below.[25]

A *CER* that yields zero profitability represents a good benchmark because all values above this imply positive social profitability. I therefore use the terms "break-even" and "benchmark" interchangeably in referring to the flexible budget benchmarks. Table 4 shows a range of discounted flexible budget *CER* benchmarks. The sensitivity analysis considers a range of values for the achievement effect on wages. The table is based on the average annual worker compensation in 2007, according to the Bureau of Labor Statistics (2008) estimates of estimated wage and fringe benefits and hours per week.

The fact that Tennessee STAR, and some other interventions, appear to meet the social profitability criterion does not mean, however, that they should be adopted. Indeed, it is important to consider the relationship between the two sets of benchmarks. There are four

possible scenarios for a given *CER* estimate: (a) the *CER* meets neither the flexible nor fixed budget benchmarks; (b) it meets the fixed budget benchmarks, but not the flexible budget benchmarks; (c) it meets the flexible fixed budget benchmarks, but not the fixed budget benchmarks; and (d) it meets both benchmarks. The uninteresting cases are (a), in which case the policy could not be recommended, and (d), where the policy could be recommended. In the case of situations (b) and (c) the appropriate decision depends on the decision-making context. A school principal or district administrator has little choice but to use all available resources, so that only the fixed budget benchmarks are relevant. Conversely, a state or federal policymaker has more budgetary authority and should require that *both* benchmarks are reached. That is, someone with budgetary authority should be cautious in adopting a policy that fails the fixed budget benchmark even if it surpasses the flexible budget benchmarks because, under the assumptions in the previous section, failing to meet the fixed budget benchmark necessarily means that some other option is available that would raise achievement even more.

Conclusion: Rethinking Effect Sizes

This study proposes an approach for interpreting effect sizes that incorporates costs and therefore facilitates a policy-relevant interpretations of educational research. By comparing the ratio of effects-to-costs for alternative interventions, it is possible to draw general conclusions and create benchmarks that aid researchers in interpreting results about individual interventions—breaking the apparent "catch-22"—and guiding policymakers towards the decisions that will help them maximize student achievement and other important student outcomes.

There are several directions for future research that can make the benchmarks a reality. First, and most obviously, it is essential that future research studies report cost data and provide cost analysis. In addition to reporting the resources used in interventions in more detail, it is important that the information reflect as closely as possible the resource units experienced by the sample of students. Second, more research is needed to test the assumptions and the prevalence and level of decay/compounding. The discussion here also relies on implicit assumptions about meaning of effect sizes (Hill et al., 2007) and the validity of the *CER*s in the sense described by Shadish, Cook & Campbell (2002) with regard to effects. Future studies should apply their validity concept to costs and *CER*s as well.[26]

If cost information were made more readily available and the assumptions were further reinforced, then it would be possible to create a reasonably complete, though still far from perfect, set of benchmarks: far benchmarks and near benchmarks, fixed budget benchmarks and flexible budget benchmarks, and benchmarks that cover all the combinations of the dimensions of substitutability. Doing so is important and directly related to a wide range of current government efforts to improve education. The U.S. Department of Education's What Works Clearinghouse is summarizing evidence about the effects of various types of interventions and there are apparently plans to consider and compare costs. Also, the federal *No Child Left Behind* law has created the ambitious goal of getting all students to proficiency by 2014. The efficient use of resources is no doubt one part of the strategy necessary to help reach that and other important education goals.

Developing such benchmarks will no doubt require considerable effort. The most recent benchmarks report in health research (DCPP, 2006), for example, has nine editors, 73 chapters, and more than 400 chapter authors. However, given the enormous resources that

go into conducting educational research, and the apparent goal of this research to improve education outcomes, such an effort seems a small price to pay in order to make the results truly useful. Educational policy should be cost-effective and so too should educational research.

References

Barnett, W.S. (1985). Benefit-cost analysis of the Perry Preschool Program and its policy implications. *Educational Evaluation and Policy Analysis*, 7(4), 333-342.

Barnett, W.S. (1992). Benefits of compensatory preschool education. *Journal of Human Resources*, 27(2), 279-312.

Barnett, W.S. (1996). *Lives in the balance: Age-27 benefit-cost analysis of the High/Scope Perry Preschool Program*. Ypsilanti, MI: High/Scope Press.

Borman, G.D. & Hewes, M. (2002). The long-term effectiveness and cost-effectiveness of Success for All. *Educational Evaluation and Policy Analysis*, 24(4), 243-266.

Borman, G.D., Hewes, G.M., Overman, L.T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, 73, 125-230.

Borman, G.D., Slavin, R., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2006). *Final reading outcomes of the national randomized field trial of Success for All*. Downloaded June 10, 2007 from http://successforall.com/index.htm.

Brewer, D.J., Krop, C., Gill, B.P., & Reichardt, R. (1999). Estimating the cost of national class size reductions under different policy alternatives, *Educational Evaluation and Policy Analysis*, 21(2), 179-192.

Briggs,A.H., O'Brien, B.J., & Blackhouse, G. (2002). Thinking outside the box: Recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. *Annual Review of Public Health*, 23, 377–401.

Browning, E.K. (1987) On the Marginal Welfare Cost of Taxation, American Economic Review, 77(1), 11-23.

Bureau of Labor Statistics (2007). *Current population survey, detailed tables*. Downloaded May 30, 2007 from http://www.bls.gov/data/home.htm.

Bureau of Labor Statistics (2008). *Current employment statistics*. Downloaded May 30, 2007 from http://www.bls.gov/data/home.htm.

Campbell, F.A., Pungello, E.P., Miller-Johnson, S., Burchinal, M. & Ramey, C.T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology*, 37(2), 231-242.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edition. Hillsdale, NJ: Erlbaum.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 97(2), 31-47.

Gill, B., Timpane, M., Ross, K., Brewer, D. (2001). *Rhetoric versus reality: What we know and what we need to know about vouchers and charter schools*. Santa Monica: The RAND Corporation.

Glass, G.V. & Smith, M.L. (1979). Meta-analysis of research on class size and Achievement. *Educational Evaluation and Policy Analysis*, 1(1), 2-16.

Grissmer, D. (1999). Class size effects: Assessing the evidence, its policy implications, and future research agenda. *Educational Evaluation and Policy Analysis*, 21(2), 231-248.

Grissmer, D., Flanagan, A., Kawata, J.H., & Williamson, S. (2000). *Improving student achievement: What NAEP scores tell us*. Santa Monica, CA: RAND.

Hanushek, E. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis*, 21 (2), 143-163.

Harris, D.N. (2000). Optimal resource allocation. Doctoral dissertation. East Lansing, MI: Michigan State University.

Harris, D.N. (2002). Identifying optimal class sizes and teacher salaries. In H.M. Levin and P.J. McEwan (Eds.), *Cost effectiveness analysis in education*. Larchmont, NY: American Education Finance Association.

Harris, D.N. (2004). *Funding Florida's schools: Adequacy, costs, and the state constitution.* Tempe, AZ: Arizona State University Education Policy Studies Laboratory.

Harris, D.N. (2007). Class size and school size: Taking the trade-offs seriously. In F.M. Hess and T.Loveless (Eds.), *Brookings papers on education policy 2006*-2007. Washington, DC: Brookings Institution.

Haveman, R. & Wolfe, B. (1984). Schooling and economic well-being: The role of non-market effects. *The Journal of Human Resources*, 14, 377-407.

Hedges, L. & Stock, W. (1983). The effects of class size: An examination of rival hypotheses. *American Educational Research Journal*, 20(1), 63-85.

Hill, C.J., Bloom, H.S., Black, A.R., & Lipsey, M.W. (2007) *Empirical Benchmarks for Interpreting Effect Sizes in Research*. Washington, DC: MDRC.

Krueger, A.B. (2003). Economic considerations and class size. *Economic Journal*, 113, 34-63.

Krueger, A.B. & Whitmore, D.M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, 111, 1-28.

Laxminarayan, R., Chow, J., and Shahid–Salles, S.A. (2006). Intervention cost–effectiveness: Overview of main messages. Disease Control Priorities in Developing Countries, 2nd edition, (pp.35-86). New York: Oxford University Press.

Levin, H.M. (1991). Cost-effectiveness at a quarter century. In M.W. McLaughlin and D.C. Phillips (Eds.), *Evaluation and education at quarter century* (pp.189-209). Chicago: University of Chicago Press.

Levin, H.M. (2002). *The cost-effectiveness of whole school reforms*. ERIC Clearinghouse on Urban Education, Urban Diversity Series 114. New York: Teachers College, Columbia University.

Levin, H.M., Glass, G.V., & Meister, G.R. (1987). Cost-effectiveness of computer-assisted instruction. *Evaluation Review*, 11(1), 50-72.

Levin, H.M. & McEwan, P.J. (2001). *Cost-effectiveness analysis*, 2nd Edition. London: Sage Publications.

Lipscomb, J., Weinstein, M.C., & Torrance, G.W. (1996). Time preference. In M.R. Gold, L.B. Russell, J.E. Siegel, & M.C. Weinstein (Eds), *Cost-Effectiveness in Health and Medicine* (pp.214-246). New York: Oxford University Press.

Lipsey, M.W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage Publications.

Lockheed, M.E. & Hanushek, E. (1988) Improving educational efficiency in developing countries: What do we know? *Compare* 18(1), 21-37.

Masse, L.N. & Barnett, W.S. (2002). *A benefit cost analysis of the Abecedarian early childhood intervention*. New Brunswick, NJ: National Institute for Early Education Research.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A. & Ehrle, K. (1999). Evaluating the SAGE program: A pilot program in targeted pupil-teacher reduction in Wisconsin. *Educational Evaluation and Policy Analysis*, 21(2), 165-177.

Monk, D.H. & King, J. (1993). Cost analysis as a tool for education reform. In S.L. Jacobson and R. Berne (Eds.), *Reforming education: Yearbook of the American Education Finance Association*. Thousand Oaks, CA: Corwin.

Moore, M.A., Boardman, A.E., Vining, A.R., Weimer, D.L., & Greenberg, D.H. (2004). Just give me a number! Practical values for the social discount rate. *Journal of Policy Analysis and Management*, 23, 789-812.

Muennig, P. (2002). *Designing and conducting cost-effectiveness analyses in medicine and health care*. San Francisco: Jossey-Bass

Nye, B.A., Hedges, L.V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Education Evaluation and Policy Analysis*, 21(2), 127-142.

Petersen, N.S., Kolen, M.J. & Hoover, H.D. (1993). Scaling, norming and equating. In R. Linn (Ed.), *Educational measurement*. Phoenix, AZ: Oryx Press.

Peterson, P.E. & Howell, W.G. (2006). *The education gap: Vouchers and urban schools*. Washington, DC: The Brookings Institution.

Polsky, D., Glick, H.A., Willke, R., & Schulman, K. (1997). Confidence intervals for cost-effectiveness ratios: A comparison of four methods. *Journal of Health Economics,* 6(3), 243-252.

Podgursky, M. (2005). *Is teacher pay adequate?* Research Working Papers Series WP05-10. Cambridge, MA: John F. Kennedy School of Government.

Rice, J.K. (1997). Cost analysis in education: Paradox and possibility. *Educational Evaluation and Policy Analysis*, 19(4), 309-317.

Rice, J.K. (2002). Cost analysis in education policy research: A comparative analysis across fields of public policy," In Henry M. Levin and Patrick J. McEwan (Eds), *Cost-effectiveness in educational policy* (pp.21-35), Larchmont, NY: Eye on Education.

Rothstein, R. & Jacobsen, R. (2006). The goals of education. *Phi Delta Kappan*, 88(4), 264-272.

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Schiefelbain, E., Wolff, L., & Scheifelbain (1999). Cost effectiveness of primary education policies in Latin America: A survey of expert opinion. *UNESCO Bulletin*, 49, 53-76.

Taylor, L. L. & Fowler, W.J. (2006). *A comparable wage approach to geographic cost adjustment*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

U.S. Department of Education, National Center on Education Statistics, Projections of through 2015 (September, 2006). Downloaded on May 30, 2007 from http://nces.ed.gov/programs/projections/tables/table_36.asp.

U.S. Census (2005). *Public education finances*. Washington, DC: Census Bureau.

Valentine, J. C. & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

Weiss, A. (1995). Human capital versus signaling theories of wages. *Journal of Economic Literature*, 9(4), 133-154.

World Bank (1993) World Development Report. Washington, DC: Author.

Table 1a: Discount Factors for Cost-Effectiveness Ratios (CERs) from K-12 Interventions (3 percent discount rate)

Age (Grade) that intervention begins	*Length of Intervention (in years)*												
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
17 (Grade 12)	25.37												
16 (Grade 11)	24.61	24.98											
15 (Grade 10)	23.87	24.23	24.60										
14 (Grade 9)	23.15	23.51	23.86	24.22									
13 (Grade 8)	22.46	22.80	23.15	23.50	23.85								
12 (Grade 7)	21.79	22.12	22.45	22.79	23.13	23.48							
11 (Grade 6)	21.13	21.45	21.78	22.11	22.44	22.77	23.11						
10 (Grade 5)	20.50	20.81	21.13	21.44	21.77	22.09	22.42	22.75					
9 (Grade 4)	19.88	20.19	20.49	20.80	21.11	21.43	21.74	22.07	22.39				
8 (Grade 3)	19.29	19.58	19.88	20.18	20.48	20.78	21.09	21.40	21.72	22.04			
7 (Grade 2)	18.71	18.99	19.28	19.57	19.86	20.16	20.46	20.76	21.07	21.37	21.68		
6 (Grade 1)	18.15	18.42	18.70	18.98	19.27	19.56	19.85	20.14	20.43	20.73	21.03	21.34	
5 (Kinderg.)	17.60	17.87	18.14	18.41	18.69	18.97	19.25	19.53	19.82	20.11	20.40	20.70	21.00
4 (Pre-K1)	17.07	17.33	17.60	17.86	18.13	18.40	18.67	18.95	19.23	19.51	19.79	20.08	20.37
3 (Pre-K2)	16.56	16.81	17.07	17.33	17.59	17.85	18.11	18.38	18.65	18.92	19.20	19.47	19.75
2	16.06	16.31	16.56	16.81	17.06	17.31	17.57	17.83	18.09	18.35	18.62	18.89	19.16
1	15.58	15.82	16.06	16.30	16.55	16.79	17.04	17.29	17.55	17.80	18.06	18.32	18.59
0	15.12	15.35	15.58	15.81	16.05	16.29	16.53	16.78	17.02	17.27	17.52	17.77	18.03

Notes: The discount factors were calculated based on: $T \cdot \left[\left(\sum_{t=18-a}^{65-a} (1-\delta)^t \right) / \left(\sum_{t=1}^{T} (1-\delta)^t \right) \right]$. This is identical to the second term of equation (3), except that the term is multiplied by the number of time periods for the intervention (*T*). This last step is necessary because the undiscounted *CER* in equation (2) is based on the sum of annual costs across years (C) whereas the costs in equation (3) are necessarily based on annual costs ($\overline{C}$). A discount rate of 0.03 is assumed in this table. See table 1b for a rate of 0.07. The correspondence between the age and grade are based on the typical student age at the beginning of the school year.

Table 1b: Discount Factors for Cost-Effectiveness Ratios (CERs) from K-12 Interventions (7 percent discount rate)

Age (Grade) that intervention begins	*Length of Intervention (in years)*												
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
17 (Grade 12)	13.81												
16 (Grade 11)	12.85	13.31											
15 (Grade 10)	11.95	12.38	12.82										
14 (Grade 9)	11.11	11.51	11.93	12.35									
13 (Grade 8)	10.33	10.71	11.09	11.48	11.89								
12 (Grade 7)	9.61	9.96	10.32	10.68	11.05	11.43							
11 (Grade 6)	8.94	9.26	9.59	9.93	10.28	10.63	11.00						
10 (Grade 5)	8.31	8.61	8.92	9.24	9.56	9.89	10.23	10.57					
9 (Grade 4)	7.73	8.01	8.30	8.59	8.89	9.20	9.51	9.83	10.15				
8 (Grade 3)	7.19	7.45	7.72	7.99	8.27	8.55	8.84	9.14	9.44	9.75			
7 (Grade 2)	6.69	6.93	7.18	7.43	7.69	7.95	8.23	8.50	8.78	9.07	9.36		
6 (Grade 1)	6.22	6.44	6.67	6.91	7.15	7.40	7.65	7.91	8.17	8.43	8.71	8.98	
5 (Kinderg.)	5.78	5.99	6.21	6.43	6.65	6.88	7.11	7.35	7.60	7.84	8.10	8.35	8.62
4 (Pre-K1)	5.38	5.57	5.77	5.98	6.19	6.40	6.62	6.84	7.06	7.30	7.53	7.77	8.01
3 (Pre-K2)	5.00	5.18	5.37	5.56	5.75	5.95	6.15	6.36	6.57	6.78	7.00	7.23	7.45
2	4.65	4.82	4.99	5.17	5.35	5.53	5.72	5.91	6.11	6.31	6.51	6.72	6.93
1	4.33	4.48	4.64	4.81	4.98	5.15	5.32	5.50	5.68	5.87	6.06	6.25	6.45
0	4.02	4.17	4.32	4.47	4.63	4.79	4.95	5.12	5.28	5.46	5.63	5.81	5.99

Notes: See notes to Table 1a, which is identical except that the discount rate here is 0.07 instead of 0.03.

Table 2: Summary of Cost-Effectiveness Ratios from Sample Empirical Studies

Intervention	*Start grade; length of intervention*	*Undiscounted CERs (d/$1,000)*			*Discounted CERs (δ=0.03)*		
		Short-term	*Medium-term*	*Long-term*	*Short-term*	*Medium-term*	*Long-term*
Class size (STAR)	K; 4 years	0.026-0.086	0.023-0.083	0.019-0.075	0.479-1.583	0.417-1.528	0.350-1.381
Computer-Aided Instr.	2,5; 1 year	0.600			11.57		
Cross-Age Tutoring							
Peer & adult	2-6; 1 year	0.900			17.89		
Peer only	2-3; 1 year	1.400			26.18		
Adult only	4-6; 1 year	0.300			6.150		
Early Childhood							
Perry Preschool	Pre-K; 2 years			0.022-0.027			0.381-0.468
Abecedarian child	Pre-K; 3 years	0.012-0.017		0.007-0.009	0.211-0.299		0.123-0.158
Instructional time	2,5; 1 year	0.400			7.716		
Success for All	K; 3 years	0.030-0.116		0.022-0.055	0.544-2.104		0.399-0.998

Notes: Sources for the undiscounted *CER*s are discussed in the text. Short-term effects are measured immediately after an intervention. Medium-term effects are measured 1-4 years after the intervention. Long-term effects are measured 5+ years after the intervention. Undiscounted *CER*s are discounted based on discount factors in Table 1a (δ=0.03). The specific discount factor used in each case can be found using the starting age and length of intervention information in the second column and using this to find the appropriate discount factor in Table 1a. In cases where effects were averaged across grades (e.g., instructional time for grades 2 and 5), the discount factor is chosen based on average of the years with downward rounding. Multiplying the undiscounted *CER*s by discount factors in Table 1a yields the present discounted value (PDV) of the *CER*s, shown in the right-hand columns. All costs in the *CER*s are based on 2007 dollars.

Table 3: Hypothetical Cost-Effectiveness Ratio Benchmarks

Benchmark	*Discounted CERs*		
	Short-term	*Medium-term*	*Long-term*
Largest *CERs* from Table 2	26.18	1.528	1.381
Largest *CER*s w/ extrapolation	26.18	19.64	13.09
Hypothetical Benchmarks			
Small	<0.4	<0.3	<0.2
Medium	0.4-6.0	0.3-4.5	0.2-3.0
Large	>6.0	>4.5	>3.0

Notes: Hypothetical benchmarks based on sample of discounted *CER*s reported in Table 2. Extrapolation based on 50 percent decay between short-term and long-term. See text for further explanation.

Table 4: Break-Even CERs based on Flexible Budget (Monetary Benefits) Approach

Effect of Achievement on Wages (γ)	*Break-Even Undiscounted CERs (d/$1,000)*			*Break-Even Discounted CERs*
	δ=0.00	*δ=0.03*	*δ= 0.07*	
0.02	0.022	0.042	0.081	1.041
0.04	0.011	0.021	0.047	0.520
0.06	0.007	0.014	0.027	0.347
0.08	0.005	0.010	0.020	0.260
0.10	0.004	0.008	0.016	0.208
0.12	0.004	0.007	0.013	0.173
0.14	0.003	0.006	0.012	0.149

Notes: Long-term (fully decayed) effects sizes assumed. Calculations based on equation (5) and average annual wage (including fringe benefits) of $48,048 based on U.S. Bureau of Labor Statistics estimates of hours per week and total compensation per week.

Notes

[1] Lipsey (1990) later conducted a meta-analyses of effect size estimates and found that Cohen's initial benchmarks are similar to what is found by separating the effect size distribution into thirds and taking the midpoint effect size within each third. While informative, these still do not consider ignore costs.

[2] Rice's discussion starts with the identification of three difficulties in expanding the use of cost analysis: (1) "difficulties associated with the conceptualization and calculation of costs and effects"; (2) "issues associated with the identification and justification of the distribution of costs and effects across stakeholder groups"; and (3) "factors that have limited the generalizability of studies conducted" (1997, p.310). These same issues are addressed here as well, though with the goal of broadening the use of cost-effectiveness among researchers. Thus, the focus here is on making generalizations about cost-effectiveness.

[3] The assumption that personnel compensation provides a reasonable measure of opportunity requires some explanation. It is assumed that for-profit firms compensate their workers closely on each individual's contribution to production. For-profit firms will not pay a worker more than she contributes, for doing so would reduce profits. At the same time, for-profits cannot underpay the worker for risk of losing the worker to another organization. Because for-profits compete for workers with non-profits and governments, it is reasonable to use the compensation paid to governmental and non-profit workers as a measure of their opportunity cost because these workers have opportunities in for-profit firms.

[4] This is often not the case in developing nations and it is up to the researcher in these cases to decide whether the time of students should be counted. Also, even in developed nations, students from low-income families are often expected to leave school early, perhaps even earlier than the law allows, to help take care of younger siblings and other relatives.

[5] The calculation of standard errors has been discussed in the health literature. See, for example, Polsky, Glick, Willke, and Schulman (1997).

[6] For those unfamiliar with inflation adjustments, these can be made using web sites such as that of the Bureau of Labor Statistics (http://data.bls.gov/cgi-bin/cpicalc.pl) which make the calculations automatically.

[7] This is one way in which health research differs from education research. Improved health always counts no matter the age of the individual.

[8] It is also possible, as noted earlier, that what appears to be decay is really just a gradual increase in the standard deviation as students get older, which would make the effect sizes appear to decline.

[9] All of the studies cited in the text recommend a baseline discount rate of exactly three percent with the slight exception of Moore et al. (2004) who present different discount rates that depend, first, on the likelihood that the intervention "crowds out" other investments. Based on their discussion, there are two reasons that education investments do not crowd out other investments: (a) most K-12 education is funded from state and local governments that face fixed budgets and therefore fund projects primarily from taxation, which mainly reduces consumption rather than investment; and (b) international market flows mean that even large investments in education are very small in comparison to the total funds available for investments. (If there were crowding out, it would be necessary to multiple costs by the "shadow price of capital" to reflect the value of the foregone investments.)

A second issue is that education involves intergenerational transfers. Moore et al. argue that in these cases the discount rate of future generations is uncertain so that a time-declining discount rate should be chosen. However, given that the value of such transfers is difficult to measure, they are excluded in the present analysis. In the conditions of the present analysis, the Moore et al. (2004) discussion suggests a discount rate of 3.5 percent. This is quite close to the 3.0 percent figure recommended in other texts; therefore, I use the latter figure throughout.

[10] It might appear that this discount factor (47) has to be divided by the number of treatment years (see the denominator), but this is not the case because the total costs on which the undiscounted *CER*s are based have already multiplied by the number of treatment years; that is, $T \cdot \overline{C} = C$.

[11] The discount factor formula for the post-K-12 interventions is: $\left(\sum_{t=a+T}^{65-a} (1-\delta)^t \right) \Big/ \left(\sum_{t=1}^{T} (1-\delta)^t \right)$ where a is the expected age at which the average participant begins the intervention.

[12] It is not entirely clear that the long-term effects found by Krueger and Whitmore truly represent a decay of the initial larger effects. The content of the earlier tests may have differed from the college entrance exam and, if the entrance exam was less well-aligned with the curriculum as one would expect, then this change in measurement may have caused the decline in effects rather than decay.

[13] Harris (2000, 2002) discusses meta-analytic results reported by Glass and Smith (1979) and Hedges and Stock (1983), as well as research on Wisconsin SAGE. The two meta-analyses are based on a similar set of studies and reach similar conclusions. The Wisconsin SAGE program involved 5,000 students to small and large classes with average sizes quite similar to the STAR program. Another limitation of SAGE for the purpose of understanding class size effects is that it included staff development, after-school programs, and a new curriculum for students in the treatment group. While it may appear difficult to isolate a true class size effect from the other aspects of this policy, Molnar et al. (1999) found that these other programs had no significant effect after controlling for class size. Harris (2000) finds that the results of STAR are quite similar GS/HS and larger than those of SAGE.

[14] The average teacher salary does not necessarily reflect the marginal cost of hiring an additional teacher because new teachers hired are typically younger than the average and therefore earn lower salaries. However, in the long run, these newly hired teachers would have higher levels of experience. Therefore, the average teacher salary is used throughout. Also, the use of national data does not account for geographic cost differences.

[15] The cost of additional classrooms vary dramatically depending on the amount of extra space now available in existing schools and the way in which additional space might be provided (portables versus new construction). A reasonable estimate can be obtained by assuming a middle ground where all additional space is provided through portable classrooms. Harris (2004) reports that the cost of purchasing a single portable classroom is $89,789. The implied annual rental cost is roughly 10 percent of this amount or $8,979. Because this school requires 2.13 additional classrooms, this implies an additional cost $19,125 or $191 per student per year, or $764 over the four years.

[16] In Abecedarian, some infants in the control group received some other form of child care. Ninety-seven of the original 111 children remained in the sample at age 4. Cognitive ability and achievement were measured using various versions of the Wechsler intelligence tests and Woodcock-Johnson tests.

[17] Barnett (1992) does not report the long-term effects of Perry in effect size form, but as 1.2 "grade level equivalents" greater than the control group. There is no fixed relationship between grade equivalents and effect sizes nor does Barnett provide sufficient information to make the transformation for this particular test. The results in the text therefore represent an approximation based on results reported in Petersen, Kolen, and Hoover (1993) for eighth-graders in the Iowa Test of Basic Skills (ITBS). Follow-up results are available for the intervening years and at age 19, but no information is available to assess the size of these effects.

[18] Levin, Glass and Meister (1987) also discuss evidence on the costs and effects of class size. However, their results are based on the review by Glass and Smith (1979), which was already discussed as part of the Harris studies (2000, 2002).

[19] Using data gathered from surveys, Rothstein and Jacobsen (2006) report a simple weighted average of the priorities of U.S. adults, school board members, state legislators, and school superintendents. The top two priorities were: "basic academic skills in core subjects" (22 percent) and "critical thinking and problem solving" (18 percent). An additional 10 percent of respondents mentioned "preparation for skilled work." All of these are arguably aligned with achievement scores. (Regarding "skilled work," see later evidence about the relationship between achievement scores and labor market earnings.) The authors report that the responses were similar across the four surveyed groups. While they, quite reasonably, interpret this as evidence that people prefer a balanced set of priorities, the point here is only to show that academic outcomes represent the highest priority.

[20] Levin (2002) proposes separating costs by outcome, but this may be quite difficult in practice. Another alternative is to assume the effects on achievement (or other measured outcomes) are a constant proportion of the effects on non-achievement outcomes. If this is the case, then the relative cost-effectiveness of each intervention would be the same across outcome measures and failing to explicitly consider all the relevant outcomes would be irrelevant because doing so would affect all the CERs proportionally. While it seems likely that the relationship between achievement and non-achievement outcomes would vary depending on the type of intervention, the possibility that there is some relationship between them reduces the problems that might arise from a violation of the assumption.

[21] For evidence questioning the effect of vouchers, see Gill et al. (2001).

[22] Even though equity is not the objective of the decision, the evidence discussed above suggests that equity might nevertheless improve when trying to maximize the average achievement. Recall that many of the interventions considered here target disadvantaged students and, even in cases such as STAR class size where the intervention itself was not targeted, the effects were larger for disadvantaged students and targeting would be an appropriate policy response. Thus, the most cost-effective way to improve average student achievement may be targeting resources to disadvantaged students.

[23] There are arguably 15 different age levels (associated with pre-K-12), four primary subjects (mathematics, reading/language arts, science, and social studies). This yields 60 different grade-subject combinations.

[24] In earlier versions of this study, I followed the approach of Harris (2007) and also accounted for annual changes in average labor market productivity (ρ). This raises the economic return to education and, for this same reason, raises the opportunity cost of teacher time. This issue is not considered here because the effects are, again, relatively small as productivity growth rates average only about one percent per year. This also simplifies the comparison with the *CER* in equation (3). Economists also often account for the cost of obtaining the financial resources necessary to pay for programs, either through debt borrowing or taxation. See Browning (1987) for an introduction. Harris (2000) recommends a marginal cost of taxation of 25 percent (applied to all costs). The issues of changes in labor market productivity and the cost of taxation/borrowing are relevant only to flexible budget analyses. Adjustments for these factors in the fixed budget analysis cancel out in the comparison of *CER*s.

[25] There are two other factors that complicate the estimation of γ: On the other hand, these labor market benefits capture only a portion of the total benefits to society (Haveman and Wolfe, 1984), suggesting that the above values for γ are too low. On the other hand, Weiss (1995) concludes from an extensive review of evidence that a majority off the relationship between education (specifically, additional years of education) and labor market benefits are likely over-estimated.

[26] I thank an anonymous reviewer for making this suggestion.