ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms

NOVEMBER 2012

Mark W. Lipsey, Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick

NCSER 2013-3000
U.S. DEPARTMENT OF EDUCATION

Page intentionally left blank.

Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms

NOVEMBER 2012

Mark W. Lipsey
Peabody Research Institute
Vanderbilt University

Kelly Puzio
Department of Teaching and Learning
Washington State University

Cathy Yun
Vanderbilt University

Michael A. Hebert
Department of Special Education and Communication Disorders
University of Nebraska-Lincoln

Kasia Steinka-Fry
Peabody Research Institute
Vanderbilt University

Mikel W. Cole
Eugene T. Moore School of Education
Clemson University

Megan Roberts
Hearing & Speech Sciences Department
Vanderbilt University

Karen S. Anthony
Vanderbilt University

and

Matthew D. Busick
Vanderbilt University

NCSER 2013-3000
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

This report was prepared for the National Center for Special Education Research, Institute of Education Sciences under Contract ED-IES-09-C-0021.

Disclaimer

The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Command Decisions Systems & Solutions to develop a report that assists with the translation of effect size statistics into more readily interpretable forms for practitioners, policymakers, and researchers. The views expressed in this report are those of the author and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

November 2012

Alternate Formats
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

There are nine authors for this report with whom IES contracted to develop the discussion of the issues presented. Mark W. Lipsey, Cathy Yun, Kasia Steinka-Fry, Megan Roberts, Karen S. Anthony, and Matthew D. Busick are employees or graduate students at Vanderbilt University; Kelly Puzio is an employee at Washington State University; Michael A. Hebert is an employee at University of Nebraska-Lincoln; and Mikel W. Cole is an employee at Clemson University. The authors do not have financial interests that could be affected by the content in this report.

Contents

List of Tables

Table **Page**

List of Figures

Figure **Page**

Page intentionally left blank.

Introduction

The superintendent of an urban school district reads an evaluation of the effects of a vocabulary building program on the reading ability of fifth graders in which the primary outcome measure was the CAT/5 reading achievement test. The mean posttest score for the intervention sample was 718 compared to 703 for the control sample. The vocabulary building program thus increased reading ability, on average, by 15 points on the CAT/5. According to the report, this difference is statistically significant, but is this a big effect or a trivial one? Do the students who participated in the program read a lot better now, or just a little better? If they were poor readers before, is this a big enough effect to now make them proficient readers? If they were behind their peers, have they now caught up?

Knowing that this intervention produced a statistically significant positive effect is not particularly helpful to the superintendent in our story. Someone intimately familiar with the CAT/5 (California Achievement Test, 5th edition; CTB/McGraw Hill 1996) and its scoring may be able to look at these means and understand the magnitude of the effect in practical terms but, for most of us, these numbers have little inherent meaning. This situation is not unusual—the native statistical representations of the findings of studies of intervention effects often provide little insight into the practical magnitude and meaning of those effects. To communicate that important information to researchers, practitioners, and policymakers, those statistical representations must be translated into some form that makes their practical significance easier to infer. Even better would be some framework for directly assessing their practical significance.

This paper is directed to researchers who conduct and report education intervention studies. Its purpose is to stimulate and guide them to go a step beyond reporting the statistics that emerge from their analysis of the differences between experimental groups on the respective outcome variables. With what is often very minimal additional effort, those statistical representations can be translated into forms that allow their magnitude and practical significance to be more readily understood by the practitioners, policymakers, and even other researchers who are interested in the intervention that was evaluated.

Organization and Key Themes

The primary purpose of this paper is to provide suggestions to researchers about ways to present statistical findings about the effects of educational interventions that might make the nature and magnitude of those effects easier to understand. These suggestions and the related discussion are framed within the context of studies that use experimental designs to compare measured outcomes for two groups of participants, one in an intervention condition and the other in a control condition. Though this is a common and, in many ways, prototypical form for studies of intervention effects, there are other important forms. Though not addressed directly, much of what is suggested here can be applied with modest adaptation to experimental studies that compare outcomes for more than two groups or compare conditions that do not include a control (e.g., compare different interventions), and to quasi-experiments that compare outcomes for nonrandomized groups. Other kinds of intervention studies that appear in educational research are beyond the scope of this paper. Most notable among those other kinds are observational studies, e.g., multivariate

analysis of the relationship across schools between natural variation in per pupil funding and student achievement, and single case research designs such as those often used in special education contexts to investigate the effects of interventions for children with low-incidence disabilities.

The discussion in the remainder of this paper is divided into three main sections, each addressing a relatively distinct aspect of the issue. The first section examines two common, but inappropriate and misleading ways to characterize the magnitude of intervention effects. Its purpose is to caution researchers about the problems with these approaches and provide some context for consideration of better alternatives.

The second section reviews a number of ways to represent intervention effects descriptively. The focus there is on how to better communicate the nature and magnitude of the effect represented by the difference on an outcome variable between the intervention and control samples. For example, it may be possible to express that difference in terms of percentiles or the contrasting proportions of intervention and control participants scoring above a meaningful threshold value. Represented in terms such as those, the nature and magnitude of the intervention effect may be more easily understood and appreciated than when presented as means, regression coefficients, *p*-values, standard errors, and the like.

The point of departure for descriptive representations of an intervention effect is the set of statistics generated by whatever analysis the researcher uses to estimate that effect. Most relevant are the means and, for some purposes, the standard deviations on the outcome variable for the intervention and control groups. Alternatively, the point of departure might be the effect size estimate, which combines information from the group means and standard deviations and is an increasingly common and frequently recommended way to report intervention effects. However, not every analysis routine automatically generates the statistics that are most appropriate for directly deriving alternative descriptive representations or for computing the effect size statistic as an intermediate step in deriving such representations. This second section of the paper, therefore, begins with a subsection that provides advice about obtaining the basic statistics that support the various representations of intervention effects that are described in the subsections that follow it.

The third section of this paper sketches some approaches that might be used to go beyond descriptive representations to more directly reveal the practical significance of an intervention effect. To accomplish that, the observed effect must be assessed in relationship to some externally defined standard, target, or frame of reference that carries information about what constitutes practical significance in the respective intervention domain. Covered in that section are approaches that benchmark effects within such frameworks as normative growth, differences between students and schools with recognized practical significance, the effects found for other similar interventions, and cost.

Inappropriate and Misleading Characterizations of the Magnitude of Intervention Effects

Some of the most common ways to characterize the effects found in studies of educational interventions are inappropriate or misleading and thus best avoided. The statistical tests routinely applied to the difference between the means on outcome variables for intervention and control samples, for instance, yield a *p* value—the estimated probability that a difference that large would be found when, in fact, there was no difference in the population from which the samples were drawn. "Very significant" differences with, say, $p < .001$ are often trumpeted as if they were indicative of especially large and important effects, ones that are more significant than if p were only marginally significant (e.g., $p = .10$) or just conventionally significant (e.g., $p = .05$). Such interpretations are quite inappropriate. The *p*-values characterize only statistical significance, which bears no necessary relationship to practical significance or even to the statistical magnitude of the effect. Statistical significance is a function of the magnitude of the difference between the means, to be sure, but it is also heavily influenced by the sample size, the within samples variance on the outcome variable, the covariates included in the analysis, and the type of statistical test applied. None of the latter is related in any way to the magnitude or importance of the effect.

When researchers go beyond simply presenting the intervention and control group means and the *p*-value for the significance test of their difference, the most common way to represent the effect is with a standardized effect size statistic. For continuous outcome variables, this is almost always the standardized mean difference effect size—the difference between the means on an outcome variable represented in standard deviation units. For example, a 10 point difference between the intervention and control on a reading achievement test with a pooled standard deviation of 40 for those two samples is .25 standard deviation units, that is, an effect size of .25.

Standardized mean difference effect sizes are a useful way to characterize intervention effects for some purposes. This effect size metric, however, has very little more inherent meaning than the simple difference between means; it simply transforms that difference into standard deviation units. Interpreting the magnitude or practical significance of an effect size requires that it be compared with appropriate criterion values or standards that are relevant and meaningful for the nature of the outcome variable, sample, and intervention condition on which it is based. We will have more to say about effect sizes and their interpretation later. We raise this matter now only to highlight a widely used but, nonetheless, misleading standard for assessing effect sizes and, at least by implication, their practical significance.

In his landmark book on statistical power, Cohen (1977, 1988) drew on his general impression of the range of effect sizes found in social and behavioral research in order to create examples of power analysis for detecting smaller and larger effects. In that context, he dubbed .20 as "small," .50 as "medium," and .80 as "large." Ever since, these values have been widely cited as standards for assessing the magnitude of the effects found in intervention research despite Cohen's own cautions about their inappropriateness for such general use. Cohen was attempting, in an unsystematic way, to describe the distribution of effect sizes one might find if one piled up all the effect sizes on all the different outcome measures for all the different interventions

targeting individual participants that were reported across the social and behavioral sciences. At that level of generality, one could take any given effect size and say it was in the low, middle, or high range of that distribution.

The problem with Cohen's broad normative distribution for assessing effect sizes is not the idea of comparing an effect size with such norms. Later in this paper we will present some norms for effect sizes from educational interventions and suggest doing just that. The problem is that the normative distribution used as a basis for comparison must be appropriate for the outcome variables, interventions, and participant samples on which the effect size at issue is based. Cohen's broad categories of small, medium, and large are clearly not tailored to the effects of intervention studies in education, much less any specific domain of education interventions, outcomes, and samples. Using those categories to characterize effect sizes from education studies, therefore, can be quite misleading. It is rather like characterizing a child's height as small, medium, or large, not by reference to the distribution of values for children of similar age and gender, but by reference to a distribution for all vertebrate mammals.

McCartney and Rosenthal (2000), for example, have shown that in intervention areas that involve hard to change low baserate outcomes, such as the incidence of heart attacks, the most impressively large effect sizes found to date fall well below the .20 that Cohen characterized as small. Those "small" effects correspond to reducing the incidence of heart attacks by about half—an effect of enormous practical significance. Analogous examples are easily found in education. For instance, many education intervention studies investigate effects on academic performance and measure those effects with standardized reading or math achievement tests. As we show later in this paper, the effect sizes on such measures across a wide range of interventions are rarely as large as .30. By appropriate norms—that is, norms based on empirical distributions of effect sizes from comparable studies—an effect size of .25 on such outcome measures is large and an effect size of .50, which would be only "medium" on Cohen's all encompassing distribution, would be more like "huge."

In short, comparisons of effect sizes in educational research with normative distributions of effect sizes to assess whether they are small, middling, or large relative to those norms should use appropriate norms. Appropriate norms are those based on distributions of effect sizes for comparable outcome measures from comparable interventions targeted on comparable samples. Characterizing the magnitude of effect sizes relative to some other normative distribution is inappropriate and potentially misleading. The widespread indiscriminate use of Cohen's generic small, medium, and large effect size values to characterize effect sizes in domains to which his normative values do not apply is thus likewise inappropriate and misleading.

Representing Effects Descriptively

The starting point for descriptive representations of the effects of an educational intervention is the set of native statistics generated by whatever analysis scheme has been used to compare outcomes for the participants in the intervention and control conditions. Those statistics may or may not provide a valid estimate of the intervention effect. The quality of that estimate will depend on the research design, sample size, attrition, reliability of the outcome measure, and a host of other such considerations. For purposes of this discussion, we assume that the researcher begins with a credible estimate of the intervention effect and consider only alternate representations or translations of the native statistics that initially describe that effect.

A closely related alternative starting point for a descriptive representation of an intervention effect is the effect size estimate. Although the effect size statistic is not itself much easier to interpret in practical terms than the native statistics on which it is based, it is useful for other purposes. Most notably, its standardized form (i.e., representing effects in standard deviation units) allows comparison of the magnitude of effects on different outcome variables and across different studies. It is thus well worth computing and reporting in intervention studies but, for present purposes, we include it among the initial statistics for which an alternative representation would be more interpretable by most users.

In the following parts of this section of the paper, we first provide advice for configuring the native statistics generated by common analyses in a form appropriate for supporting alternate descriptive representations. We include in that discussion advice for configuring the effect size statistic as well in a few selected situations that often cause confusion.

Configuring the Initial Statistics that Describe an Intervention Effect to Support Alternative Descriptive Representations

Covariate Adjustments to the Means on the Outcome Variable

Several of the descriptive representations of intervention effects described later are derived directly from the means and perhaps the standard deviations on the outcome variable for the intervention and control groups. However, the *observed* means for the intervention and control groups may not be the best choice for representing an intervention effect. The difference between those means reflects the effect of the intervention, to be sure, but it may also reflect the influence of any initial baseline differences between the intervention and control groups. The value of random assignment to conditions, of course, is that it permits only chance differences at baseline, but this does not mean there will be no differences, especially if the samples are not large. Moreover, attrition from posttest measurement undermines the initial randomization so that estimates of effects may be based on subsets of the intervention and control samples that are not fully equivalent on their respective baseline characteristics even if the original samples were.

Researchers often attempt to adjust for such baseline differences by including the respective baseline values as covariates in the analysis. The most common and useful covariate is the pretest for the outcome measure along with basic demographic variables such as age, gender, ethnicity, socioeconomic status, and the like.

Indeed, even when there are no baseline differences to account for, the value of such covariates (especially the pretest) for increasing statistical power is so great that it is advisable to routinely include any covariates that have substantial correlations with the posttest (Rausch, Maxwell, Kelley 2003). With covariates included in the analysis, the estimation of the intervention effect is the difference between the *covariate-adjusted* means of the intervention and control samples. These adjusted values better estimate the actual intervention effect by reducing any bias from the baseline differences and thus are the best choices for use in any descriptive representation of that effect. When that representation involves the standard deviations, however, their values should not be adjusted for the influence of the covariates. In virtually all such instances, the standard deviations are used as estimates of the corresponding population standard deviations on the outcome variables without consideration for the particular covariates that may have been used in estimating the difference on the means.

When the analysis is conducted in analysis of covariance format (ANCOVA), most statistical software has an option for generating the covariate-adjusted means. When the analysis is conducted in multiple regression format, the unstandardized regression coefficient for the intervention dummy code (intervention=1, control=0; or +0.5 vs. -0.5) is the difference between the covariate-adjusted means. In education, analyses of intervention effects are often multilevel when the outcome of interest is for students or teachers who, in turn are nested within classrooms, schools, or districts. Using multilevel regression analysis, e.g., HLM, does not change the situation with regard to the estimate of the difference between the covariate-adjusted means—it is still the unstandardized regression coefficient on the intervention dummy code. The unadjusted standard deviations for the intervention and control groups, in turn, can be generated directly by most statistical programs, though that option may not be available within the ANCOVA, multiple regression, or HLM routine itself.

For binary outcomes, such as whether students are retained in grade, placed in special education status, or pass an exam, the analytic model is most often logistic regression, a specialized variant of multiple regression for binary dependent variables. The regression coefficient (β) in a logistic regression for the dummy coded variable representing the experimental condition (e.g., 1=intervention, 0=control) is a covariate-adjusted log odds ratio representing the intervention effect (Crichton 2001). Unlogging it ($\exp^{\beta}$) produces the covariate-adjusted odds ratio for the intervention effect, which can then be converted back into the terms of the original metric.

For example, an intervention designed to improve the passing rate on an algebra exam might produce the results shown below. The odds of passing for a given group are defined as the ratio of the number (or proportion) who pass to the number (or proportion) who fail. For the intervention group, therefore, the odds of passing are 45/15=3.0 and, for the control group, the odds are 30/30=1.0. The odds ratio characterizing the intervention effect is the ratio of these two values, that is 3/1=3, and indicates that the odds of passing are three times greater for a student in the intervention group than for one in the control group.

	Passed	Failed
Intervention	45	15
Control	30	30

Suppose the researcher analyzes these outcomes in a logistic regression model with race, gender, and prior math achievement scores included as covariates to control for initial differences between the two groups. If the coefficient on the intervention variable in that analysis, converted to a covariate-adjusted odds-ratio, turns out to be 2.53, it indicates that the unadjusted odds ratio overestimated the intervention effect because of baseline differences that favored the intervention group. With this information, the researcher can construct a covariate-adjusted version of the original 2x2 table that estimates the proportions of students passing in each condition when the baseline differences are taken into account. To do this, the frequencies for the control sample and the total N for the intervention sample are taken as given. We then want to know what passing frequency, p, for the intervention group allows the odds ratio, (p·30)/((60-p)·30), to equal 2.53. Solving for p reveals that it must be 43. The covariate-adjusted results, therefore, are as shown below. Described as simple percentages, the covariate-adjusted estimate is that the intervention increased the 50% pass rate of the control condition to 72% (43/60) in the intervention condition.

	Passed	Failed
Intervention	43	17
Control	30	30

Identifying or Obtaining Appropriate Effect Size Statistics

A number of the ways of representing intervention effects and assessing their practical significance described later in this paper can be derived directly from the standardized mean difference effect size statistic, commonly referred to simply as the effect size. This effect size is defined as the difference between the mean of the intervention group and the mean of the control group on a given outcome measure divided by the pooled standard deviations for those two groups, as follows:

$$ES = \frac{\overline{X}_T - \overline{X}_C}{s_p}$$

Where $\overline{X}_T$ is the mean of the intervention sample on an outcome variable, $\overline{X}_C$ is the mean of the control sample on that variable, and s_p is the pooled standard deviation. The pooled standard deviation is obtained as the square root of the weighted mean of the two variances, defined as:

$$s_p = \sqrt{\frac{(n_T - 1)s_T^2 + (n_C - 1)s_C^2}{n_T + n_C - 2}}$$

where n_T and n_C are the number of respondents in the intervention and control groups, and s_T and s_C are the respective standard deviations on the outcome variable for the intervention and control groups.

The effect size is typically reported to two decimal places and, by convention, has a positive value when the intervention group does better on the outcome measure than the control group and a negative sign when it does worse. Note that this may not be the same sign that results from subtraction of the control mean from the intervention mean. For example, if low scores represent better performance, e.g., as with a measure of the number of errors made, then subtraction will yield a negative value when the intervention group performs better than the control, but the effect size typically would be given a positive sign to indicate the better performance of the intervention group.

Effect sizes can be computed or estimated from many different kinds of statistics generated in intervention studies. Informative sources for such procedures include the What Works Clearinghouse *Procedures and Standards Handbook* (2011; Appendix B) and Lipsey and Wilson (2001; Appendix B). Here we will only highlight a few features that may help researchers identify or configure appropriate effect sizes for use in deriving alternative representations of intervention effects. Moreover, many of these features have implications for statistics other than the effect size that are involved in some representations of intervention effects.

Clear understanding of what the numerator and denominator of the standardized mean difference effect size represent will allow many common mistakes and confusions in the computation and interpretation of effect sizes to be avoided. The numerator of the effect size estimates the difference between the experimental groups on the means of the outcome variable that is attributable to the intervention. That is, the numerator should be the best estimate available of the mean intervention effect estimated in the units of the original metric. As described in the previous subsection, when researchers include baseline covariates in the analysis, the best estimate of the intervention effect is the difference between the covariate-adjusted means on the outcome variable, not the difference between the unadjusted means.

The purpose of the denominator of the effect size is to standardize the difference between the outcome means in the numerator into metric free standard deviation units. The concept of *standardization* is critical here. Standardization means that each effect size is represented in the same way, i.e., in a standard way, irrespective of the outcome construct, the way it is measured, or the way it is analyzed. The sample standard deviations used for this purpose estimate the corresponding population standard deviations on the outcome measure. As such, the standard deviations should not be adjusted by any covariates that happened to be used in the design or analysis of the particular study. Such adjustments would not have general applicability to other designs and measures and thus would compromise the standardization that is the point of representing the intervention effect in standard deviation units. This means that the raw standard deviations for the intervention and control samples should be pooled into the effect size denominator, even when multilevel analysis models with complex variance structures are used.

Pooling the sample standard deviations for the intervention and control groups is intended to provide the best possible estimate of the respective population standard deviation by using all the data available. This procedure assumes that both those standard deviations estimate a common population standard deviation. This is the homogeneity of variance assumption typically made in the statistical analysis of intervention effects. If homogeneity of variance cannot be assumed, then consideration has to be given to the reason why

the intervention and control group variances differ. In a randomized experiment, this should not occur on outcome variables unless the intervention itself affects the variance in the intervention condition. In that case, the better estimate may be the standard deviation of the control group even though it is estimated on a smaller sample than the pooled version.

In the multilevel situations common in education research, a related matter has to do with the population that is relevant for purposes of standardizing the intervention effect. Consider, for example, outcomes on an achievement test that is, or could be, used nationally. The variance for the national population of students can be partitioned into between and within components according to the different units represented at different levels. Because state education systems differ, we might first distinguish between-state and within-state variance. Within states, there would be variation between districts; within districts, there would be variation between schools; within schools, there would be variation between classrooms; and within classrooms, there would be variation between students. The total variance for the national population can thus be decomposed as follows (Hedges 2007):

$$\sigma^2_{Total} = \sigma^2_{States} + \sigma^2_{Districts} + \sigma^2_{Schools} + \sigma^2_{Classrooms} + \sigma^2_{Students}$$

In an intervention study using a national sample, the sample estimate of the standard deviation includes all these components. Any effect size computed with that standard deviation is thus standardizing the effect size with the national population variance as the reference value. The standard deviation computed in a study using a sample of students from a single classroom, on the other hand, estimates only the variance of the population of students who might be in that classroom in that school in that district in that state. In other words, this standard deviation does not include the between classroom, between school, between district, and between state components that would be included in the estimate from a national sample. Similarly, an intervention study that draws its sample from one school, or one district, will yield a standard deviation estimate that is implicitly using a narrower population as the basis for standardization than a study with a broader sample. This will not matter if there are no systematic differences on the respective outcome measure between students in different states, districts, schools, and classrooms, i.e., those variance components are zero. With student achievement measures, we know this is generally not the case (e.g., Hedges and Hedberg 2007). Less evidence is available for other measures used in education intervention studies, but it is likely that most of them also show nontrivial differences between these different units and levels.

Any researcher computing effect sizes for an intervention study or using them as a basis for alternative representations of intervention effects should be aware of this issue. Effect sizes based on samples of narrower populations will be larger than effect sizes based on broader samples even when the actual magnitudes of the intervention effects are identical. And, that difference will be carried through to any other representation of the intervention effect that is based on the effect size. Compensating for that difference, if appropriate, will require adding or subtracting estimates of the discrepant variance components, with the possibility that those components will have to be estimated from sources outside the research sample itself.

The discussion above assumes that the units on which the sample means and standard deviations are computed for an outcome variable are individuals, e.g., students. The nested data structures common in education intervention studies, however, provide different units on which means and standard deviations can be computed, e.g., students, clusters of students in classrooms, and clusters of classrooms in schools. For instance, in a study of a whole school intervention aimed at improving student achievement, with some schools assigned to the intervention condition and others to the control, there are two effect sizes the researcher could estimate. The conventional effect size would standardize the intervention effect estimated on student scores using the pooled student level standard deviations. Alternatively, the student level scores might be aggregated to the school level and the school level means could be used to compute an effect size. That effect size would represent the intervention effect in standard deviation units that reflect the variance between schools, not that between students. The result is a legitimate effect size, but the school units on which it is based make this effect size different from the more conventional effect size that is standardized on variation between individuals.

The numerators of these two effect sizes would not necessarily differ greatly. The respective means of the student scores in the intervention and control groups would be similar to the means of the school-level means for those same students unless the number of students in each school differs greatly and is correlated with the school means. However, the standard deviations will be quite different because the variance between schools is only one component of the total variance between students. Between-school variance on achievement test scores is typically around 20-25% of the total variance, the intraclass correlation coefficient (ICC) for schools (Hedges and Hedberg 2007). The between schools standard deviation thus will be about $\sqrt{.25}$ = .50 or less of the student level standard deviation and the effect size based on school units will be about twice as large as the effect size based on students as the units even though both describe the same intervention effect.

Similar situations arise in multilevel samples whenever the units on which the outcome is measured are nested within higher level clusters. Each such higher level cluster allows for its own distinctive effect size to be computed. A researcher comparing effect sizes in such situations or, more to the point for present purposes, using an effect size to derive other representations of intervention effects, must know which effect size is being used. An effect size standardized on a between-cluster variance component will nearly always be larger than the more conventional effect size standardized on the total variance across the lower level units on which the outcome was directly measured. That difference in numerical magnitude will then be carried into any alternate representation of the intervention effect based on that effect size and the results must be interpreted accordingly.

Descriptive Representations of Intervention Effects

Representation in Terms of the Original Metric

Before looking at different ways of transforming the difference between the means of the intervention and control samples into a different form, we should first consider those occasional situations in which differences on the original metric are easily understood without such manipulations. This occurs when the

units on the measure are sufficiently familiar and well defined that little further description or interpretation is needed. For example, an outcome measure for a truancy reduction program might be the proportion of days on which attendance was expected for which the student was absent. The outcome score for each student, therefore, is a simple proportion and the corresponding value for the intervention or control groups is the mean proportion of days absent for the students in that group. Common events of this sort in education that can be represented as counts or proportions include dropping out of school, being expelled or suspended, being retained in grade, being placed in special education status, scoring above a proficiency threshold on an achievement test, completing assignments, and so forth.

Intervention effects on outcome measures that involve well recognized and easily understood events can usually be readily interpreted in their native form by researchers, practitioners, and policymakers. Some caution is warranted, nevertheless, in presenting the differences between intervention and control groups in terms of the proportions of such events. Differences between proportions can have different implications depending on whether those differences are viewed in absolute or relative terms. Consider, for example, a difference of three percentage points between the intervention and control groups in the proportion suspended during the school year. Viewed in absolute terms, this appears to be a small difference. But relative to the suspension rate for the control sample a three point decrease might be substantial. If the suspension rate for the control sample is only 5%, for instance, a decrease of three percentage points reduces that rate by more than half. On the other hand, if the control sample has a suspension rate of 40%, a reduction of three percentage points might rightly be viewed as rather modest.

In some contexts, the numerical values on an outcome measure that does not represent familiar events may still be sufficiently familiar that differences are well-understood despite having little inherent meaning. This might be the case, for instance, with widely used standardized tests. For example, the Peabody Picture Vocabulary Test (PPVT; Dunn and Dunn 2007), one of the most widely used tests in education, is normed so that standard scores have a mean of 100 for the general population of children at any given age. Many researchers and educators have sufficient experience with this test to understand what scores lower or higher than 100 indicate about children's skill level and how much of an increase constitutes a meaningful improvement. Generally speaking, however, such familiarity with the scoring of a particular measure of this sort is not widespread and most audiences will need more information to be able to interpret intervention effects expressed in terms of the values generated by an outcome measure.

Intervention Effects in Relation to Pre-Post Change

When pretest measures of an outcome variable are available, the pretest means may be used to provide an especially informative representation of intervention effects using the original metric. This follows from the fact that the intent of interventions is to bring about change in the outcome; that is, change between pretest and posttest. The full representation of the intervention effect, therefore, is not simply the difference between the intervention and control samples on the outcome measure at posttest, but the differential change between pretest and posttest on that outcome. By showing effects as differential change, the researcher reveals not only the end result but the patterns of improvement or decline that characterize the intervention and control groups.

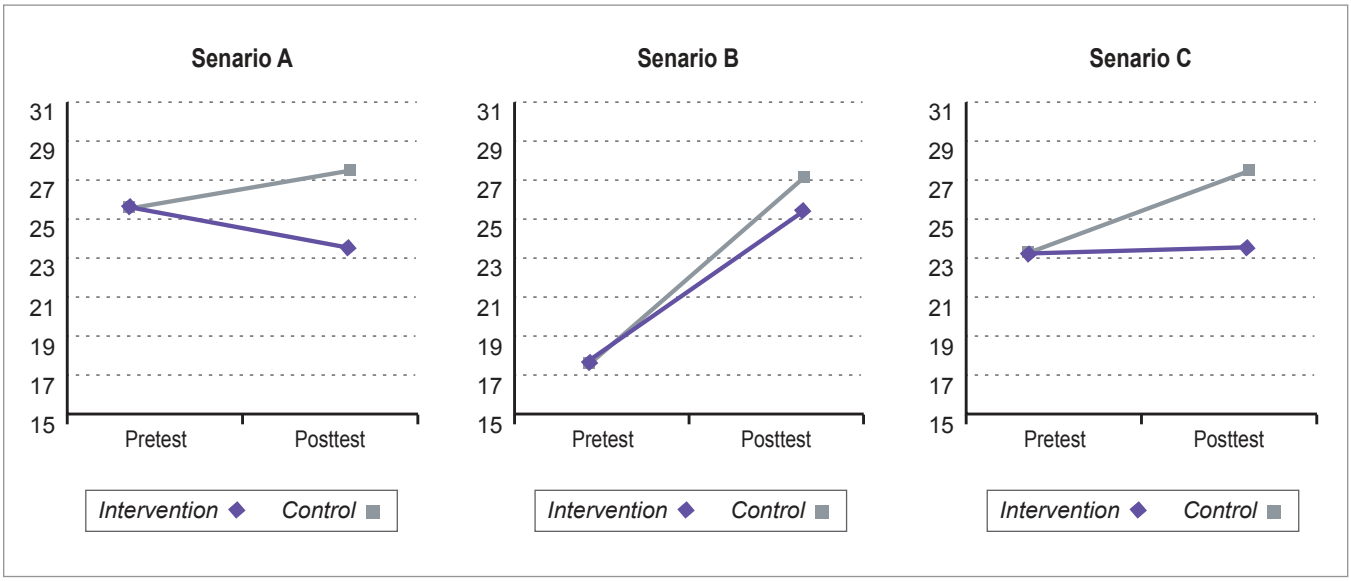
Consider, for example, a program that instructs middle school students in conflict resolution techniques with the objective of decreasing interpersonal aggression. Student surveys at the beginning and end of the school year are administered for intervention and control schools that provide composite scores for the amount of physical, verbal, and relational aggression students experience. These surveys show significantly lower levels for the intervention schools than the control schools, indicating a positive effect of the conflict resolution program, say a mean of 23.8 for the overall total score for the students in the intervention schools and 27.4 for the students in the control schools. That 3.6 point favorable outcome difference, however, could have come from any of a number of different patterns of change over the school year for the intervention and control schools. Table 1 below shows some of the possibilities, all of which assume an effective randomization so that the pretest values at the beginning of the school year were virtually identical for the intervention and control schools.

Table 1. Pre-post change differentials that result in the same posttest difference

	Scenario A		**Scenario B**		**Scenario C**	
	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Intervention	25.5	23.8	17.7	23.8	22.9	23.8
Control	25.6	27.4	17.6	27.4	23.0	27.4

As can be seen even more clearly in Figure 1, for Scenario A the aggression levels decreased somewhat in the intervention schools while increasing in the control schools. In Scenario B, the aggression levels increased quite a bit (at least relative to the intervention effect) in both samples, but the amount of the increase was not as great in the intervention schools as the control schools. In Scenario C, on the other hand, there was little change in the reported level of aggression over the course of the year in the intervention schools, but things got much worse during that time in the control schools. These different patterns of differential pre-post change depict different trajectories for aggression absent intervention and give different pictures of what it is that the intervention accomplished. In Scenario A it reversed the trend that would have otherwise occurred. In Scenario B, it ameliorated an adverse trend, but did not prevent it from getting worse. In Scenario C, the intervention did not produce appreciable improvement over time, but kept the amount of aggression from getting worse.

Figure 1. Pre-post change for the three scenarios with the same posttest difference

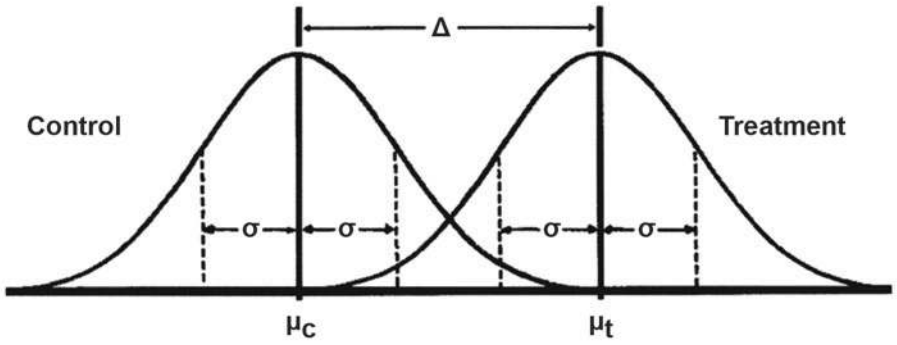


As this example illustrates, a much fuller picture of the intervention effect is provided when the difference between the intervention and control samples on the outcome variable is presented in relation to where those samples started at the pretest baseline. A finer point can be put on the differential change for the intervention and control groups, if desired, by proportioning the intervention effect against the control group pre-post change. In Scenario B above, for instance, the difference between the control group's pretest and posttest composite aggression scores is 9.8 (27.4 - 17.6) while the posttest difference between the intervention and control group is -3.6 (23.8 - 27.4). The intervention, therefore, reduced the pre-post increase in the aggression score by 36.7% (-3.6/9.8).

Overlap Between Intervention and Control Distributions

If the distributions of scores on an outcome variable were plotted separately for the intervention and control samples, they might look something like Figure 2 below. The magnitude of the intervention effect is represented directly by the difference between the means of the two distributions. The standardized mean difference effect size, discussed earlier, also represents the difference between the two means, but does so in standard deviation units. Still another way to represent the difference between the outcomes for the intervention and control groups is in terms of the overlap between their respective distributions. When the difference between the means is larger, the overlap is smaller; when the difference between the means is smaller, the overlap is larger. The amount of overlap, in turn, can be described in terms of the proportion of individuals in each distribution who are above or below a specified reference point on one of the distributions. Proportions of this sort are often easy to understand and appraise and, therefore, may help communicate the magnitude of the effect. Various ways to take advantage of this circumstance in the presentation of intervention effects are described below.

Figure 2. Intervention and control distributions on an outcome variable



Notation in this figure: μ_c and μ_t are the means of the control and treatment distribution respectively; Δ is the difference between those means; and σ is the common standard deviation of both distributions.

Intervention Effects Represented in Percentile Form

For outcomes assessed with standardized measures that have norms, a simple representation of the intervention effect is to characterize the means of the control and intervention samples according to the percentile values they represent in the norming distribution. For a normed, standardized measure, these values are often provided as part of the scoring scheme for that measure. On a standardized math reading achievement test, for instance, suppose that the mean of the control sample fell at the 47th percentile according to the test norms for the respective age group and the mean of the intervention sample fell at the 52nd percentile. This tells us, first, that the mean outcome for the control sample is somewhat below average performance (50th percentile) relative to the norms and, second, that the effect of the intervention was to improve performance to the point where it was slightly above average. In addition, we see that the corresponding increase was 5 percentile points. That 5 percentile point difference indicates that the individuals receiving the intervention, on average, have now caught up with the 5% of the norming population that otherwise scored just above them.

In their study of Teach for America, Decker, Mayer, and Glazerman (2004) used percentiles in this way to characterize the statistically significant effect they found on student math achievement scores. Decker et al. also reported the pretest means as percentiles so that the relative gain of the intervention sample was evident. This representation revealed that the students in the control sample were at the 15th percentile at both the pretest and posttest whereas the intervention sample gained 3 percentiles by moving from the 14th to the 17th percentile.

It should be noted that the percentile differences on the norming distribution that are associated with a given difference in scores will vary according to where the scores fall in the distribution. Table 2 shows the percentile levels for the mean score of the lower scoring experimental group (e.g., control group when its mean score is lower than that of the treatment group) in the first column. The numbers in the body of the table then show the corresponding percentile level of the other group (e.g., treatment) that are associated

with a range of score differences represented in standard deviation units (which therefore means these differences can also be interpreted as standardized mean difference effect sizes). As shown there, if one group scores at the 50th percentile and the other has a mean score that is .50 standard deviations higher, that higher group will be at the 69th percentile for a difference of 19 in the percentile ranking. A .50 standard deviation difference between a group scoring at the 5th percentile and a higher scoring group will put that other group at the 13th percentile for a difference of only 8 in the percentile ranking. This same pattern for differences between two groups also applies to pre-post gains for one group. Researchers should therefore be aware that intervention effects represented as percentile differences or gains on a normative distribution can look quite different depending on whether the respective scores fall closer to the middle or the extremes of the distribution.

Table 2. Upper percentiles for selected differences or gains from a lower percentile

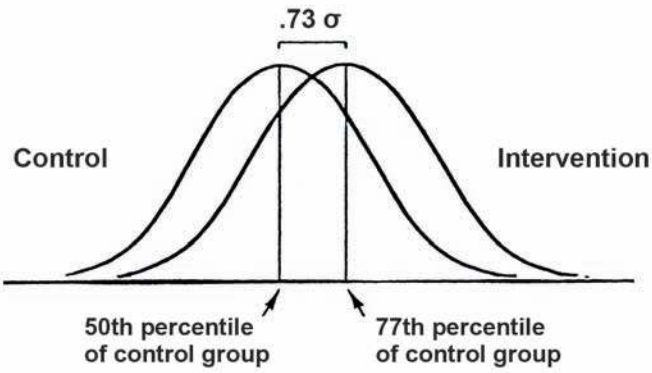
Lower Percentile	Difference or Gain in Standard Deviations				
	.10	.20	.50	.80	1.00
5th	6	7	13	20	26
10th	12	14	22	32	39
15th	17	20	30	41	48
25th	28	32	43	54	62
50th	54	58	69	79	84
75th	78	81	88	93	95
85th	87	89	94	97	98
90th	92	93	96	98	99
95th	96	97	98	99	99

NOTE: Table adapted from Albanese (2000).

A similar use of percentiles can be applied to the outcome scores in the intervention and control groups when those scores are not referenced to a norming distribution. The distribution of scores for the control group, which represents the situation in the absence of any influence from the intervention under study, can play the role of the norming distribution in this application. The proportion of scores falling below the control group and intervention group means can then be transformed into the corresponding percentile values on the control distribution. These values can be obtained from the cumulative frequency tables that most statistical analysis computer programs readily produce for the values on any variable. For a symmetrical distribution, the mean of the control sample will be at the 50th percentile (the median). The mean score for the intervention sample can then be represented in terms of its percentile value on that same control distribution. Thus we may find that the mean for the intervention group falls at the 77th percentile of the control distribution, indicating that its mean is now higher than 77% of the scores in the control sample. With a control group mean at the 50th percentile, another way of describing the difference is that the intervention has moved 27% of the sample from a score below the control mean to one above that mean.

This comparison is shown in Figure 3.

Figure 3. Percentile values on the control distribution of the means of the control and intervention groups



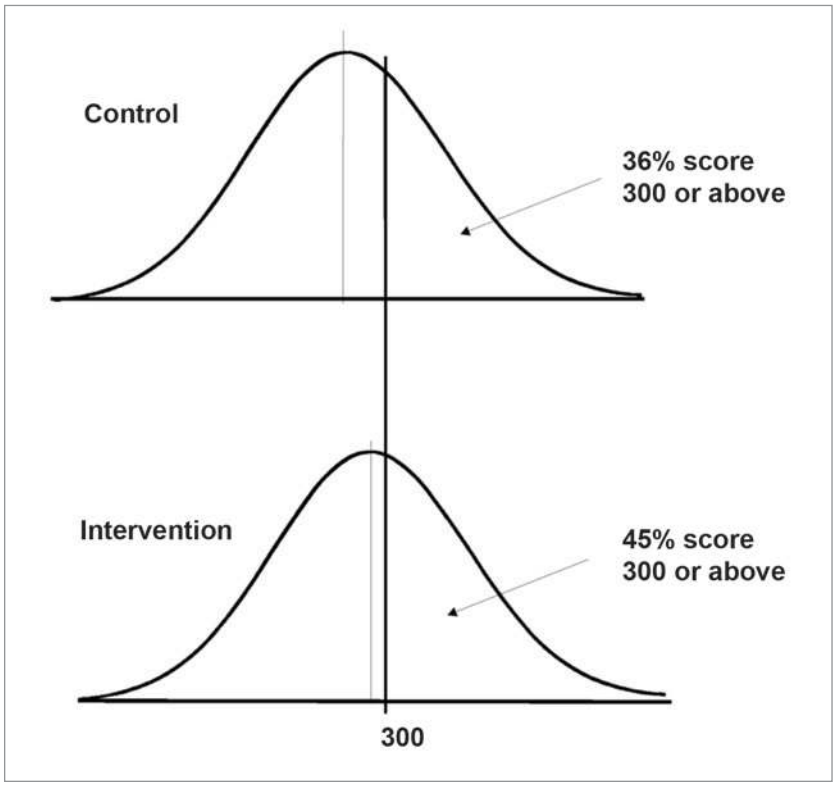
NOTE: Adapted from Redfield and Rousseau 1981. The common standard deviation of the two distributions is represented by σ.

Intervention Effects Represented as Proportions Above or Below a Reference Value

The representation of an intervention effect as the percentiles on a reference distribution, as described above, is based on the proportions of the respective groups above and below a specific threshold value on that reference distribution. A useful variant of this approach is to select an informative threshold value on the control distribution and depict the intervention effect in terms of the proportion of intervention cases above or below that value in comparison to the corresponding proportions of control cases. The result then indicates how many more of the intervention cases are in the desirable range defined by that threshold than expected without the intervention.

When available, the most meaningful threshold value for comparing proportions of intervention and control cases is one externally defined to have substantive meaning in the intervention context. Such threshold values are often defined for criterion-referenced tests. For example, thresholds have been set for the National Assessment of Educational Progress (NAEP) achievement tests with cutoff scores that designate *Basic*, *Proficient*, and *Advanced* levels of performance. On the NAEP math achievement test, for instance, scores between 299 and 333 are identified as indicating that 8th grade students are proficient. If we imagine that we might assess a math intervention using the NAEP test, we could compare the proportion of students in the intervention versus control conditions who scored 300 or above—that is, were at least minimally proficient. Figure 4 shows what the results might look like. In this example, 36% of the control students scored above that threshold level whereas 45% of the intervention students did so.

Figure 4. Proportion of the control and intervention distributions scoring above an externally defined proficiency threshold score



Similar thresholds might be available from the norming data for a standardized measure. For example, the mean standard score for the Peabody Picture Vocabulary test (PPVT; Dunn and Dunn 2007) is 100, which is, therefore, the mean age-adjusted score in the norming sample. Assuming representative norms, that score represents the population average for children of any given age. For an intervention with the PPVT as an outcome measure, the intervention effect could be described in terms of the proportion of children in the intervention versus control samples scoring 100 or above. If the proportion for either sample is at least .50, it tells us that their performance is average for their age. Suppose that for a control group, 32% scored 100 or above at posttest, identifying them as a low performing sample. If 38% of the intervention group scored 100 or above, we see that the effect of the intervention has been to move 6% of the children from the below average to the above average range. At the same time, we see that this has not been sufficient to close the gap between them and normative performance on this measure.

With a little effort, researchers may be able to identify meaningful threshold values for measures that do not already have one defined. Consider a multi-item scale on which teachers rate the problem behavior of students in their classrooms. When pretest data are collected on this measure, the researcher might also ask each teacher to nominate several children who are borderline—not presenting significant behavior problems, but close to that point. The scores of those children could then be used to identify the approximate point on the rating scale at which teachers begin to view the classroom behavior of a child as problematic. That score then provides a threshold value that allows the researcher to describe the effects of, say, a classroom behavior management program in terms of how many fewer students in the intervention condition than the control

condition fall in the problem range. Differences on the means of an arbitrarily scored multi-item rating scale, though critical to the statistical analysis, are not likely to convey the magnitude of the effect as graphically as this translation into proportions of children above a threshold teachers themselves identify.

Absent a substantively meaningful threshold value, an informative representation of the intervention effect might still be provided with a generic threshold value. Cohen (1988), for instance, used the control group mean as a general threshold value to create an index he called U3, one of several indices he proposed to describe the degree of non-overlap between control and intervention distributions. The example shown in Figure 3, presented earlier to illustrate the use of percentile values, similarly made the control group mean the key reference value.

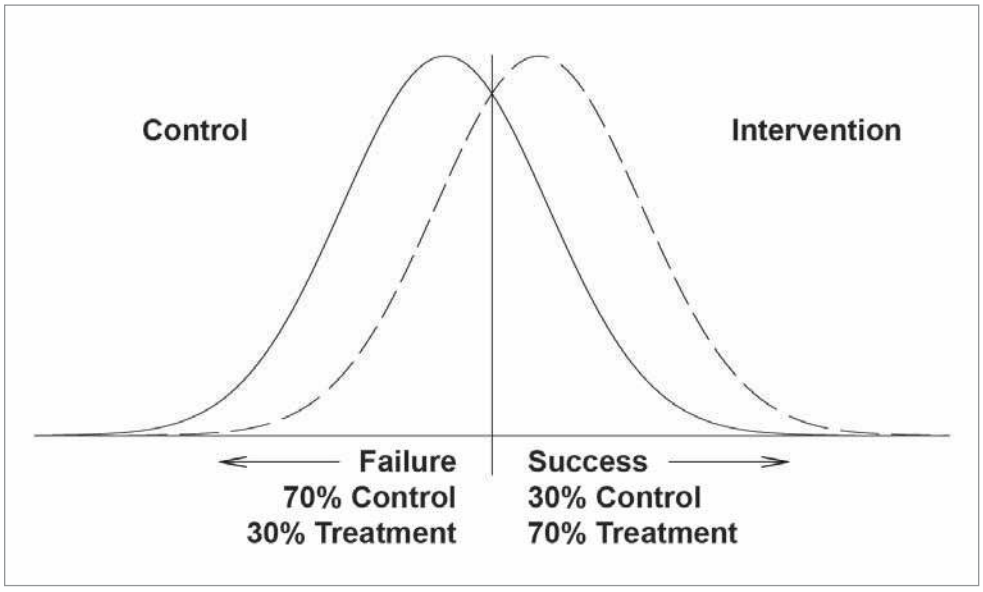
With the actual scores in hand for the control and intervention groups, it is straightforward for a researcher to determine the proportion of each above (or below) the control mean. Assuming normal distributions, those proportions and the corresponding percentiles for the control and intervention means can easily be linked to the standardized mean difference effect size through a table of areas under the normal curve. The mean of a normally distributed control sample is at the 50th percentile with a z-score of zero. Adding the standardized mean difference effect size to that z-score then identifies the z-score of the intervention mean on the control distribution. With a table of areas under the normal curve, that z-score, in turn, can be converted to the equivalent percentile and proportions in the control distribution. Table 3 shows the proportion of intervention cases above the control sample mean for different standardized mean difference effect size values, assuming normal distributions—Cohen's (1988) U3 index. In each case, the increase over .50 indicates the additional proportion of the cases that the intervention has pushed above that control condition mean.

Rosenthal and Rubin (1982) described yet another generic threshold for comparing the relative proportions of the control and intervention groups attaining it within a framework they called the Binomial Effect Size Display (BESD). In this scheme, the key success threshold value is the grand median of the combined intervention and control distributions. When there is no intervention effect, the means of both the intervention and control distributions fall at that grand median. As the intervention effect gets larger and the intervention and control distributions separate, smaller proportions of the control distribution and larger proportions of the intervention distribution fall above that grand median. Figure 5 depicts this situation.

Table 3. Proportion of intervention cases above the mean of the control distribution

Effective Size	Proportion above the Control Mean	Effect Size	Proportion above the Control Mean
.10	.54	1.30	.90
.20	.58	1.40	.92
.30	.62	1.50	.93
.40	.66	1.60	.95
.50	.69	1.70	.96
.60	.73	1.80	.96
.70	.76	1.90	.97
.80	.79	2.00	.98
.90	.82	2.10	.98
1.00	.84	2.20	.99
1.10	.86	2.30	.99
1.20	.88	2.40	.99

Figure 5. Binomial effect size display—Proportion of cases above and below the grand median



Using the grand median as the threshold value makes the proportion of the intervention sample above the threshold value equal to the proportion of the control sample below that value. The difference between these proportions, which Rosenthal and Rubin called the BESD Index, indicates how many more intervention cases are above the grand median than control cases. Assuming normal distributions, the BESD can also be linked to the standardized mean difference effect size. An additional and occasionally convenient feature of the BESD is that it is equal to the effect size expressed as a correlation; that is, the correlation between the treatment variable (coded as 1 vs. 0) and the outcome variable. Many researchers are more familiar with

correlations than standardized mean differences, so the magnitude of the effect expressed as a correlation may be somewhat more interpretable for them. Table 4 shows the proportions above and below the grand median and the BESD as the intervention effect sizes get larger. It also shows the corresponding correlational equivalents for each effect size and BESD.

Table 4. Relationship of the effect size and correlation coefficient to the BESD

Effective Size	*r*	Proportion of control/ intervention cases above the grand median	BESD (difference between the proportions)
.10	.05	.47 / .52	.05
.20	.10	.45 / .55	.10
.30	.15	.42 / .57	.15
.40	.20	.40 / .60	.20
.50	.24	.38 / .62	.24
.60	.29	.35 / .64	.29
.70	.33	.33 / .66	.33
.80	.37	.31 / .68	.37
.90	.41	.29 / .70	.41
1.00	.45	.27 / .72	.45
1.10	.48	.26 / .74	.48
1.20	.51	.24 / .75	.51
1.30	.54	.23 / .77	.54
1.40	.57	.21 / .78	.57
1.50	.60	.20 / .80	.60
1.60	.62	.19 / .81	.62
1.70	.65	.17 / .82	.65
1.80	.67	.16 / .83	.67
1.90	.69	.15 / .84	.69
2.00	.71	.14 / .85	.71
2.10	.72	.14 / .86	.72
2.20	.74	.13 / .87	.74
2.30	.75	.12 / .87	.75
2.40	.77	.11 / .88	.77

All the variations on representing the proportions of the intervention and control group distributions above or below a threshold value require dichotomizing the respective distributions of scores. It should be noted that we are not advocating that the statistical analysis be conducted on any such dichotomized data. It is well known that such crude dichotomizations discard useful data and generally weaken the analysis (Cohen 1983; MacCallum et al. 2002). What is being suggested is that, after the formal statistical analysis

is done and the results are known, depicting intervention effects in one of the ways described here may communicate their magnitude and practical implications better than means, standard deviations, t-test values, and the other native statistics that result directly from the analysis.

In applying any of these techniques, some consideration should be given to the shape of the respective distributions. When the outcome scores are normally distributed, the application of these techniques is relatively tidy and straightforward. When the data are not normally distributed, the respective empirical distributions can always be dichotomized to determine what proportions of cases are above or below any given reference value of interest, but the linkage between those proportions and other representations may be problematic or misleading. Percentile values, and differences in those values, for instance, take on quite a different character in skewed distributions with long tails than in normal distributions, as do the standard deviation units in which standardized mean difference effect sizes are represented.

Standard Scores and Normal Curve Equivalents (NCE)

Standard scores are a conversion of the raw scores on a norm referenced test that draws upon the norming sample used by the test developer to characterize the distribution of scores expected from the population for which the test is intended. A linear transform of the raw scores is applied to produce tidier numbers for the mean and standard deviation. For many standardized measures, for instance, the standard score mean may be set at 100 with a standard deviation of 15.

Presenting intervention effects in terms of standard scores can make those effects easier to understand in some regards. For example, the mean scores for the intervention and control groups can be easily assessed in relation to the mean for the norming sample. Mean scores below the standardized mean score, e.g., 100, indicate that the sample, on average, scores below the mean for the population represented in the norming sample. Similarly, a standard score mean of, say, 95 for the control group and 102 for the intervention group indicates that the effect of the intervention was to improve the scores of an underperforming group to the point where their scores were more typical of the average performance of the norming sample.

An important characteristic of standard scores for tests and measures used to assess student performance is that those scores are typically adjusted for the age of the respective students. The population represented in the norming sample from which the standard scores are derived is divided into age or school grade groups and the standard scores are determined for each group. Thus the standard scores for, say, the students in the norming sample who are in the fourth grade and average 9 years of age may be scaled to have a mean of 100 and a standard deviation of 15, but so will the standard scores for the students in the sixth grade with an average age of 11 years. Different standardized measures may use different age groups for this purpose, e.g., differing by as little as a month or two or as much as a year or more.

These age adjustments of standard scores have implications for interpreting changes in those scores over time because those changes are depicted relative to the change for same aged groups in the norming sample. A control sample with a mean standard score of 87 on the pretest and a mean score of 87 on the posttest a year later has not failed to make gains but, rather, has simply kept abreast of the differences by age in the norming sample. On the other hand, an intervention group with a mean pretest standard score of 87 and mean

posttest score of 95 has improved at a rate faster than that represented in the comparable age differences in the norming sample. This characteristic allows some interpretation of the extent to which intervention effects accelerate growth, though that depends heavily on the assumption that the sample used in the intervention study is representative of the norming sample used by the test developer.

Reporting intervention effects in standard score units thus has some modest advantages for interpretability because of the implicit comparison with the performance of the norming sample. Moreover, the means and standard deviations for standard scores are usually assigned simple round numbers that are easy to remember when making such comparisons. In other ways standard scores are not so tidy. Most notably, standard scores typically have a rather odd range. With a normal distribution encompassing more than 99% of the scores within ± 3 standard deviations, standard scores with a mean of 100 and a standard deviation of 15 will range from about 55 at the lowest to about 145 at the highest. These are not especially intuitive numbers for the bottom and top of a measurement scale. For this reason, researchers may prefer to represent treatment effects in terms of some variant of standard scores. One such variant that is well known in education is the normal curve equivalent.

Normal curve equivalents. Normal curve equivalents (NCE) are a metric developed in 1976 for the U.S. Department of Education for reporting scores on norm-referenced tests and allowing comparison across tests (Hills 1984; Tallmadge and Wood 1976). NCE scores are standard scores based on an alternative scaling of the z-scores for measured values in a normal distribution derived from the norming sample for the measure. Unlike the typical standard score, as described above, NCE scores are scaled so that they range from a low around 0 to a high of around 100, with a mean of 50. NCE scores, therefore, allow scores, differences in scores, and changes in scores to be appraised on a 100 point scale that starts at zero.

NCE scores are computed by first transforming the original raw scores into normalized z-scores. The z-score is the original score minus the mean for all the scores divided by the standard deviation; it indicates the number of standard deviations above or below a mean of zero that the score represents. The NCE score is then computed as NCE = 21.06(z-score) + 50; that is, 21.06 times the z-score plus 50. This produces a set of NCE scores with a mean of 50 and a standard deviation of 21.06. Note that the standard deviation for NCE scores is not as tidy as the round number typically used for other standard scores, but it is required to produce the other desirable characteristics of NCE scores.

As a standard score, NCEs are comparable across all the measures that derive and provide NCE scores from their norming samples if those samples represent the same population. Thus while a raw score of 82 on a particular reading test would not be directly comparable to the same numerical score on a different reading test measuring the same construct but scaled in a different way (i.e., a different mean and standard deviation), the corresponding NCE scores could be compared. For example, if the NCE score corresponding to 82 on the first measure was 68 and that corresponding to 82 on the second measure was 56, we could rightly judge that the first student's reading performance was better than that of the second student.

When NCE scores are available or can be derived from the scoring scheme for a normed measure, using them to report the pretest and posttest means for the intervention and control samples may help readers better understand the nature of the effects. It is easier to judge the difference between the intervention and control means when the scores are on a 0-100 scale than when they are represented in a less intuitive metric. Thus a 5-point difference on a 0-100 scale might be easier to interpret than a 5-point difference on a raw score metric that ranges from, e.g., 143 to 240. NCE scores also preserve the advantage of standard scores described above of allowing implicit comparisons with the performance of the norming sample. Thus mean scores over 50 show better performance than the comparable norming sample and mean scores under 50 show poorer performance.

Although standard scores, and NCE scores in particular, offer a number of advantages as a metric with which to describe intervention effects, they have several limitations. First, standard scores are all derived from the norming sample obtained by the developer of the measure. Thus these scores assume that sample is representative of the population of interest to the intervention study and that the samples in the study, in turn, are representative of the norming sample. These assumptions could easily be false for intervention studies that focus on populations distinctly appropriate for the intervention of interest. Similar discrepancies could arise for any age-adjusted standard score if the norming measures and the intervention measures were administered at very different times during the school year—differences could then be the result of predictable growth over the course of that year (Hills 1984).

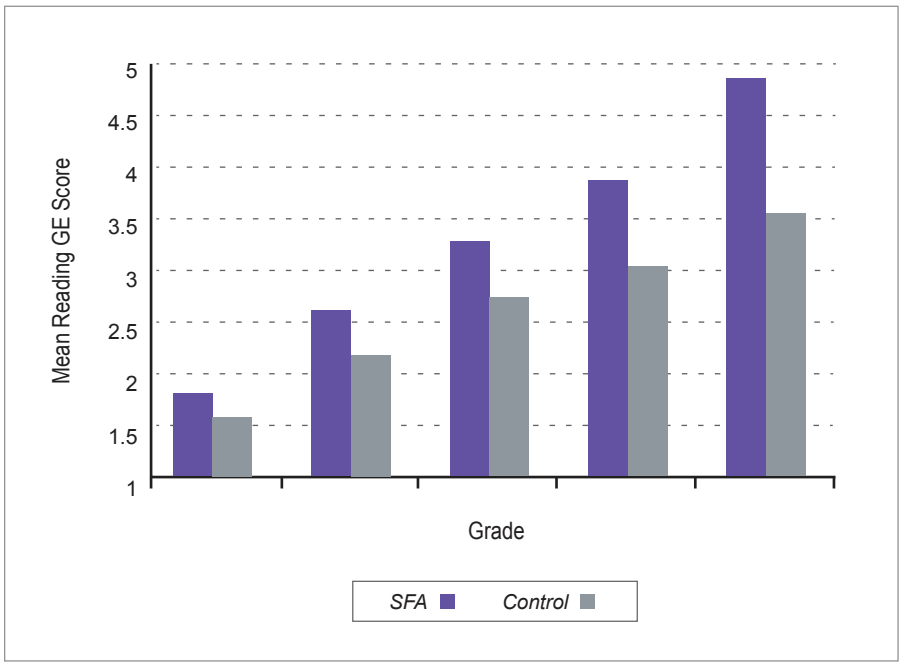
Grade Equivalent Scores

A grade equivalent (GE) is a developmental score reported for many norm-referenced tests that characterizes students' achievement in terms of the grade level of the students in the normative sample with similar performance on that test. Grade equivalent scores are based on the nine-month school year and are represented in terms of the grade level and number of full months within a nine-month school year. A GE score thus corresponds to the mean level of performance at a certain point in time in the school year for a given grade. The grade level is represented by the first number in the GE score and the month of the school year follows after a period with months ranging from a value of 0 (September) to 9 (June). A GE of 6.2, for example, represents the score that would be achieved by an average student in the sixth grade after completion of the second full month of school. The difference between GE scores of 5.2 (November of grade 5) and 6.2 (November of grade 6) represents one calendar year's growth or change in performance.

The GE score for an individual student in a given grade, or the mean for a sample of students, is inherently comparative. A GE score that differs from the grade level a student is actually in indicates performance better or worse than that of the average students at that same grade level in the norming sample. If the mean GE for a sample of students tested near the end of the fourth grade is 5.3, for instance, these students are performing at the average level of students tested in December of the fifth grade in the norming sample; that is, they are performing better than expected for their actual grade level. Conversely, if their mean GE is 4.1, they are performing below what is expected for their actual grade level. These comparisons, of course, assume that the norming sample is representative of the population from which the research sample is drawn.

Intervention effects are represented in terms of GE scores simply as the difference between the intervention and control sample means expressed in GE units. For example, in a study of Success for All (SFA), a comprehensive reading program, Slavin and colleagues (1996) compared the mean GE scores for the intervention and control samples on a reading measure used as a key outcome variable. Though the numerical differences in mean GE scores between the samples were not reported, Figure 6 shows the approximate magnitudes of those differences. The fourth grade students in SFA, for instance, scored on average about 1.8 GE ahead of the control sample, indicating that their performance was closer to that of the mean for fourth graders in the norming sample than that of the control group. Note also that the mean GE for each sample taken by itself identifies the group's performance level relative to the normative sample. The fourth grade control sample mean, at 3.0, indicates that, on average, these students were not performing up to the grade level mean in the norming sample whereas the SFA sample, by comparison, was almost to grade level.

Figure 6. Mean reading grade equivalent (GE) scores of success for all and control samples [Adapted from Slavin et al. 1996]



The GE score is often used to communicate with educators and parents because of its simplicity and inherent meaningfulness. It makes performance relative to the norm and the magnitude of intervention effects easy to understand. Furthermore, when used to index change over time, the GE score is an intuitive way to represent growth in a student's achievement. The simplicity of the GE score is deceptive, however, and makes it easy for teachers, school administrators, parents, and even researchers to misinterpret or misuse it (Linn and Slinde 1977; Ramos 1996). The limitations and issues surrounding the GE score must be understood for it to be used appropriately to represent intervention effects.

For example, as grade level increases, the difference between mean achievement in one grade and mean achievement in the next grade gets smaller. Thus one kindergarten month and one twelfth grade month are different in terms of how much average achievement growth occurs. That is, one GE score unit corresponds

to different numbers of raw score points (e.g., number of questions answered correctly) depending on the grade level. Differences in GE scores used to characterize intervention effects, therefore, can be misleading with regard to the amount of achievement gain actually produced by the intervention. An intervention effect of 0.5 GE for first grade students represents a much larger absolute increase in achievement than an intervention effect of 0.5 GE for tenth grade students even though the GE equivalents are the same. GE scores thus are not very appropriate for quantifying student growth or comparing student achievement across grade level boundaries. The apparent magnitude of intervention effects, and even the direction of those effects, can be influenced by this inconsistency in GE units across grade levels (Roderick and Nagaoka 2005).

Within a grade level, GE scores have a different problem. In conjunction with the decrease in mean differences between grades, within-grade variance tends to increase with grade level, so there will be less variance in achievement among students in the first month of kindergarten than among students in the first month of twelfth grade. With unequal within-grade variances, the same number of GE units corresponds to different magnitudes of actual achievement in different grades. A third grader who is 2.0 GE units below grade level, for instance, might be in the 1st percentile for third graders whereas a ninth grader who scores 2.0 GE units below grade level might be in the 39th percentile for ninth grade students. Despite the equal GE lags, the third grader's relative performance is much poorer than the ninth grader's. When used to characterize intervention effects, therefore, an intervention that brings the scores of targeted third graders from two GE below grade level up to grade level has brought about much more relative change than an intervention that does the same thing for ninth graders starting two GE below grade level.

In addition to their psychometric limitations, there are interpretive pitfalls when using GE scores. A common misunderstanding stems from ignoring their normative nature and interpreting GE as benchmarks that all students should be expected to achieve. From that perspective it might be supposed that all students should be on grade level, that is, have GE scores that match their actual grade level. However, normative scores such as GE inherently have a distribution that includes below and above average scores. A GE score of 6.5 for a student in February of the sixth grade year does not mean that the student has met some benchmark for being on grade level, it only means that the student's score matched the mean score of the students in the norming sample who took the test in February of their sixth grade year. It is expected that approximately half of the sixth grade students will score lower than this average and half will score higher. Similarly, a GE score of 7.5 does not necessarily mean that student is capable of doing 7th grade work but, rather, that the student shows above average performance for a 6th grader. Thus, students with GE scores below the mean are not necessarily in need of intervention and students above the mean are not necessarily ready for promotion to a higher grade. Nor is it appropriate to assume that the goal of an intervention should be to bring every student performing "below grade-level" up to grade level.

Assessing the Practical Significance of Intervention Effects

In the previous sections we have described various ways to present intervention effects that might make their nature and magnitude easier to comprehend than their native statistical form of means, regression coefficients, *p*-values, and the like. For the most part those techniques are simply translations or conversions of the statistical results. Such techniques may make it easier for a reader to understand the numerical results by representing them in a more familiar idiom, but they leave largely unanswered the question of the practical significance of the effect. Practical significance is not an inherent characteristic of the numbers and statistics that result from intervention research—it is something that must be judged in some context of application. To interpret the practical significance of an intervention effect, therefore, it is necessary to invoke an appropriate frame of reference external to its statistical representation. We must have benchmarks that mark off degrees of recognized practical or substantive significance against which we can assess the intervention effect. There are many substantive frames of reference that can provide benchmarks for this purpose, and no one will be best for every intervention circumstance.

Below we describe a number of useful benchmarks for assessing the practical significance of the effects of educational interventions. We focus on student achievement outcomes, but analogous benchmarks can be created for other outcomes. These benchmarks represent four complementary perspectives whereby intervention effects are assessed (a) relative to normal student academic growth, (b) relative to policy-relevant gaps in student performance, (c) relative to the size of the effects found in prior educational interventions, and (d) relative to the costs and benefits of the intervention. Benchmarks from the first perspective will answer questions like: How large is the effect of a given intervention if we think about it in terms of what it might add to a year of average academic growth for the target population of students? Benchmarks from the second perspective will answer questions like: How large is the effect if we think about it in terms of its ability to narrow a policy-relevant gap in student performance? Benchmarks from the third perspective will answer questions like: How large is the effect if we think about it in relation to what prior interventions have been able to accomplish? Benchmarks from the fourth perspective will answer questions like: Do the benefits of a given intervention outweigh its costs?

Benchmarking Against Normative Expectations for Academic Growth

This benchmark compares the effects of educational interventions to the natural growth in academic achievement that occurs during a year of life for an average student. Our discussion draws heavily on prior work reported in Bloom, Hill, Black, and Lipsey (2008), which can be consulted for further details. These authors computed standardized mean difference effect sizes for year-to-year growth from national norming studies for standardized tests of reading, math, science, and social studies. Those data show the growth in average student achievement from one year to the next, growth that reflects the effects of attending school plus the many other developmental influences students experience during a year of life. If we represent an intervention effect on achievement outcomes as an effect size, we can compare it with the achievement

effect sizes for annual growth and judge how much the intervention would accelerate that growth. That comparison provides one perspective on the practical significance of the intervention effect.

Table 5 reports the mean effect sizes for annual grade-to-grade reading, math, science, and social studies gains based on information from nationally-normed, longitudinally scaled achievement tests. These effect sizes are computed as the difference between the means for students tested in the spring of one school year and students in the next highest grade also tested in the spring of the school year. They thus estimate the increases in achievement that have taken place from grade to grade.

The first column of Table 5 lists mean effect size estimates for seven reading achievement tests (only six for K-1). Note the striking sequence of effect size values—annual student growth is by far the greatest during the early grades and declines thereafter. For example, the mean effect size for the period between first and second grade is 0.97; between grades five and six it is 0.32 and between grades eight and nine it is 0.24. There are a few exceptions to this pattern in some of the individual tests contributing to these means, but the overall trend is rather consistent across all of them. A similar sequence characterizes the year-to-year achievement gains in math, science, and social studies.

Although the basic patterns of the developmental trajectories are similar for all four academic subjects, the mean effect sizes for particular grade-to-grade differences vary noticeably. For example, the grade 1-2 difference shows mean annual gains for reading and math (effect sizes of 0.97 and 1.03) that are markedly higher than the gains for science and social studies (0.58 and 0.63). From about the sixth grade onward, the gains are similar across subject areas.

These year-to-year gains for average students nationally describe normative growth on standardized achievement tests that can provide benchmarks for interpreting the practical significance of intervention effects. The effect sizes on similar achievement measures for interventions with students in a given grade can be compared with the effect size representation of the annual gain expected for students at that grade level. This is a meaningful comparison when the intervention effect can be viewed as adding to students' gains beyond what would have occurred during that year without the intervention.

Table 5. Annual achievement gain: Mean effect sizes across seven nationally-normed tests

Grade Transition	Reading	Math	Science	Social Studies
Grade K - 1	1.52	1.14	--	--
Grade 1 - 2	0.97	1.03	0.58	0.63
Grade 2 - 3	0.60	0.89	0.48	0.51
Grade 3 - 4	0.36	0.52	0.37	0.33
Grade 4 - 5	0.40	0.56	0.40	0.35
Grade 5 - 6	0.32	0.41	0.27	0.32
Grade 6 - 7	0.23	0.30	0.28	0.27
Grade 7 - 8	0.26	0.32	0.26	0.25
Grade 8 - 9	0.24	0.22	0.22	0.18
Grade 9 - 10	0.19	0.25	0.19	0.19
Grade 10 - 11	0.19	0.14	0.15	0.15
Grade 11 - 12	0.06	0.01	0.04	0.04

NOTES: Adapted from Bloom, Hill, Black, and Lipsey (2008). Spring-to-spring differences are shown. The means shown are the simple (unweighted) means of the effect sizes from all or a subset of seven tests: CAT5, SAT9, Terra Nova-CTBS, Gates-MacGinitie, MAT8, Terra Nova-CAT, and SAT10.

For example, Table 5 shows that students gain about 0.60 standard deviations on nationally normed standardized reading achievement tests between the spring of second-grade and the spring of third-grade. Suppose a reading intervention is targeted to all second-graders and studied with a practice-as-usual control group of second-graders who do not receive the intervention. An effect size of, say, 0.15 on reading achievement scores for that intervention will, therefore, represent about a 25 percent improvement over the annual gain otherwise expected for second-graders—an effect quite likely to be judged to have practical significance in an elementary school context. That same effect size for tenth graders, on the other hand, would nearly double their annual gain and, by that benchmark, would likely be viewed as a rather stupendous effect.

Though we have illustrated the concept of benchmarking intervention effects against year-to-year growth with national norms for standardized achievement tests, the most meaningful comparisons will be with annual gain effect sizes on the outcome measures of interest for the particular population to which the intervention is directed. Such data will often be available from school records for prior years, or it may be collected by the researcher from nonintervention samples specifically for this purpose. Nor is this technique limited to achievement outcomes. Given a source of annual data, similar benchmarks can be obtained for any educational outcome measure that shows growth over time.

Benchmarking Against Policy-Relevant Performance Gaps

Often educational interventions are aimed at students who are performing below norms or expectations. The aspiration of those interventions is to close the gap between those students and their higher performing counterparts or, at least, to appreciably reduce the gap. In such circumstances, a straightforward basis for interpreting the practical significance of the intervention effect is to compare it with the magnitude of that gap. In this section we describe several different kinds of gaps on standardized achievement test scores that might be appropriate benchmarks against which to appraise intervention effects. This discussion also draws on prior work reported in Bloom, Hill, Black, and Lipsey (2008), where further details can be found. Though these examples deal with achievement outcomes, analogous benchmarks could be developed for any outcome variable on which there were policy-relevant gaps between the target population and an appropriate comparison population.

Benchmarking Against Differences among Students

Many educational interventions are explicitly or implicitly intended to reduce achievement gaps between different student demographic groups, for example, Black and White students or economically disadvantaged and more advantaged students. Building on the work of Konstantopoulos and Hedges (2008), we can represent the typical magnitude of such gaps as standardized mean effect sizes to which the effect size from an intervention study can be compared. This approach will be illustrated first using information from the National Assessment of Educational Progress (NAEP) and then with standardized test scores in reading and math from a large urban school district.

Performance gaps in NAEP scores. To calculate an effect size for a performance gap between two groups requires the means and standard deviations of their respective scores. For instance, published data from the 2002 NAEP indicate that the national average fourth-grade scaled reading test score is 198.75 for African-American students and 228.56 for White students. The difference in means (-29.81) divided by the standard deviation (36.05) for all fourth-graders yields an effect size of -0.83. Therefore, an effect size of, say, .20 for an intervention that improved the reading scores of fourth-grade African-American students on an achievement test similar to the NAEP could be judged in terms of the practical significance of reducing the Black-White gap by about one-fourth.

Table 6 reports standardized mean difference effect sizes for the performance differences between selected groups of students on the NAEP reading and math tests. The first panel presents effect sizes for reading. At every grade level African-American students, on average, have lower reading scores than White students with corresponding effect sizes of 0.83 for fourth-graders, decreasing to .67 for students in the 12th grade. The next two columns show a similar pattern for the gap between Hispanic students and White students, and for the gap between students who are and are not eligible for the free or reduced-price lunch programs. These latter gaps are smaller than the Black-White gap but display the same pattern of decreasing magnitude with increasing grade level, though the smaller gap in the 12th grade may reflect differential dropout rates rather than a narrowing of actual achievement differences. The second panel in Table 6 presents effect size estimates for the corresponding gaps in math performance. Expressed as effect sizes, these gaps are larger than the gaps for reading but also show a general pattern of decreases across grade levels.

Table 6. Demographic performance gaps on mean NAEP scores as effect sizes

Subject and Grade	Black-White	Hispanic-White	Eligible-Ineligible for Free/ Reduced Price Lunch
Reading			
Grade 4	-0.83	-0.77	-0.74
Grade 8	-0.80	-0.76	-0.66
Grade 12	-0.67	-0.53	-0.45
Math			
Grade 4	-0.99	-0.85	-0.85
Grade 8	-1.04	-0.82	-0.80
Grade 12	-0.94	-0.68	-0.72

NOTE: Adapted from Bloom, Hill, Black, and Lipsey (2008).
SOURCES: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Reading Assessment and 2000 Mathematics Assessment.

Konstantopoulos and Hedges (2008) found similar patterns among high school seniors using 1996 long-term trend data from NAEP and those gaps are well documented elsewhere. Performance differences such as these are a matter of considerable concern at national and local levels. For an intervention targeting the lower performing group in any of these comparisons, such gaps therefore provide a natural benchmark for assessing the practical magnitude of its effects.

Performance gaps in an urban school district. The preceding example with a nationally representative sample of students would not necessarily apply to the performance gaps in any given school district. For purposes of benchmarking intervention effects, it would be most appropriate, if possible, to use data from the context of the intervention to create these effect size benchmarks. To illustrate this point, Table 7 shows the group differences that were reported by Bloom, Hill, Black, and Lipsey (2008) for reading and math scores on the Stanford Achievement Test, 10th Edition (SAT 10; Harcourt Assessment, 2004) in a large urban school district.

The first panel in Table 7 presents effect sizes for reading. In this school district the Black-White and Hispanic-White reading achievement gaps were a full standard deviation or more and tended to increase, rather than decrease, across the grades. Relative to the race/ethnicity gaps, the gaps between students eligible and ineligible for free or reduced-price lunch programs were smaller, though still larger than for the national data shown in Table 6. The second panel in Table 7 shows the analogous effect sizes for math. These effect sizes also are generally larger than the effect sizes found in the NAEP sample for the race/ethnicity gaps, but not for the free and reduced price lunch differences.

The findings in Tables 6 and 7 illustrate a number of points about benchmarks for assessing intervention effects based on policy-relevant gaps in student performance. First, the effect size for a particular intervention, say, 0.15 on a standardized achievement test, would constitute a smaller substantive change

relative to some gaps (e.g., Black-White) than for others (e.g., socioeconomic status indexed by free or reduced price lunch eligibility). Thus, it is important to interpret an intervention's effect in the context of its target group. A second implication that can be drawn from these examples is that policy-relevant gaps for demographic subgroups may differ for achievement in different subject areas (here, reading and math) and for different grades (here, grades 4, 8, and 11 or 12). Thus, when interpreting an intervention effect in relation to a policy-relevant gap, it is important to make the comparison for the relevant outcome measure and grade level. Third, benchmarks derived from local sources in the context of the intervention will provide more relevant, and potentially different, benchmarks than those derived from national data.

Table 7. Demographic performance gaps on SAT 9 scores in a large urban school district as effect sizes

Subject and Grade	Black-White	Hispanic-White	Eligible-Ineligible for Free/ Reduced Price Lunch
Reading			
Grade 4	-1.09	-1.03	-0.86
Grade 8	-1.02	-1.14	-0.68
Grade 12	-1.11	-1.16	-0.58
Math			
Grade 4	-0.95	-0.71	-0.68
Grade 8	-1.11	-1.07	-0.58
Grade 12	-1.20	-1.12	-0.51

NOTES: Adapted from Bloom, Hill, Black, and Lipsey (2008). District local outcomes are based on SAT-9 scaled scores for tests administered in spring 2000, 2001, and 2002. SAT 9: Stanford Achievement Tests, 9th Edition (Harcourt Educational Measurement, 1996).

Benchmarking Against Differences among Schools

For school-level interventions, such as whole school reform initiatives or other school wide programs, the intervention effects might be compared with the differences on the outcome measure between poor performing and average schools. This can be thought of as a school-level gap which, if closed, would boost the performance of the weak school into a normative average range. To illustrate the construction of such benchmarks, Bloom, Hill, Black, and Lipsey (2008) adapted a technique Konstantopoulos and Hedges (2008) applied to national data to estimate what the difference in achievement would be if an average school and a weak school in the same district were working with comparable students (i.e. those students with the same demographic characteristics and past performance). Average schools were defined as schools at the 50th percentile of the school performance distribution in a given district and weak schools were defined as schools at the 10th percentile.

These school-level achievement gaps were represented as effect sizes standardized on the respective student-level standard deviations. The mean scores for 10th and 50th percentile schools in the effect size numerator were estimated from the distribution across schools of regression-adjusted mean student test scores. Those means were adjusted for prior year test scores (reading or math) and student demographic characteristics

(see Bloom et al. 2008 for details). The objective of these statistical adjustments was to estimate what the mean student achievement score would be for each school if they all had students with similar background characteristics and prior test performance.

Table 8 lists the effect sizes for the school-level performance gaps between weak and average schools in four school districts for which data were available. Although these estimates vary across grades, districts, and academic subject, almost all of them are between 0.20 and 0.40. As benchmarks for assessing the effects of whole school interventions intended to impact the achievement of the entire school or, perhaps, an entire grade within a school, these values are rather striking for their modest size. For example, if an intervention were to improve achievement for the total student body in a school or grade by an effect size of 0.20, it would be equivalent to closing half, or even all, of the performance gap between weak and average schools. This conclusion is consistent with that of Konstantopoulos and Hedges (2008) from their analysis of data from a national sample of students and schools. When viewed in this light, such an intervention effect, which otherwise might seem quite modest, would almost certainly be judged to be of great practical significance. Of course, it does not follow that it would necessarily be easy to have such an effect on an entire student body, but these benchmarks do put into perspective what might constitute a meaningful school-level effect.

Table 8. Performance gaps between average and weak schools as effect sizes

	School District			
	A	**B**	**C**	**D**
Reading				
Grade 3	0.31	0.18	0.16	0.43
Grade 5	0.41	0.18	0.35	0.31
Grade 7	0.25	0.11	0.30	NA
Grade 10	0.07	0.11	NA	NA
Math				
Grade 3	0.29	0.25	0.19	0.41
Grade 5	0.27	0.23	0.36	0.26
Grade 7	0.20	0.15	0.23	NA
Grade 10	0.14	0.17	NA	NA

NOTES: Adapted from Bloom, Hill, Black, and Lipsey (2008). "NA" indicates that a value is not available due to missing test score data. Means are regression-adjusted for test scores in prior grade and students' demographic characteristics. The tests are the ITBS for District A, SAT9 for District B, MAT for District C, and SAT8 for District D.

It is important to emphasize that the effect size estimates for weak versus average school achievement gaps reported here assume that the students in the schools compared have equal prior achievement scores and background characteristics. This assumption focuses on school effects net of variation across schools in student characteristics. The actual differences between low and average performing schools, of course, result from factors associated with the characteristics of the students as well as factors associated with school

effectiveness. The policy relevant performance gaps associated with student characteristics were discussed above with regard to differences among student demographic groups. Which of these gaps, or combinations of gaps, are most relevant for interpreting an intervention effect will depend on the intent of the intervention and whether it primarily aims to change school performance or student performance.

Benchmarking Against the Observed Effect Sizes for Similar Interventions

Another basis for assessing the practical significance of an intervention effect is to compare it with the effects found for similar interventions with similar research samples and outcome measures. This constitutes a kind of normative comparison—an effect is judged relative to the distribution of effects found in other studies. If that effect is among the largest produced by any similar intervention on that outcome with that target population, it has practical significance by virtue of being among the largest effects anyone has yet demonstrated. Conversely, if it falls in an average or below average range on that distribution, it is not so impressive relative to the current state of the art and may have diminished practical significance.

Focusing again on standardized mean difference effect sizes for achievement outcomes, we have assembled data that provide examples of such normative distributions. A search was made for published and unpublished research reports dated 1995 or later that investigated the effects of an educational intervention on achievement outcomes for mainstream K-12 students in the U.S. Low performing and at-risk student samples were included, but not specialized groups such as special education students, English language learners, or clinical samples. To ensure that the effect sizes extracted from these reports were relatively good indications of actual intervention effects, studies were restricted to those using random assignment designs with practice-as-usual control groups and attrition rates no higher than 20%. Effect sizes for the achievement outcomes and other descriptive information were ultimately coded from 124 studies that provided 181 independent samples and 829 achievement effect sizes.

The achievement effect sizes from this collection of studies can be arrayed in different ways that provide insight into what constitutes a large or small effect relative to one or another normative distribution. One distinctive pattern is a large difference in the average effect sizes found for different kinds of achievement test measures. We have categorized these measures as (a) standardized tests that cover a broad subject matter (such as the SAT 9 composite reading test), (b) standardized tests that focus on a narrower topic (such as the SAT 9 vocabulary test), and (c) specialized tests developed specifically for an intervention (such as a reading comprehension measure developed by the researcher for text similar to that used in the intervention). Table 9 presents these findings by grade level (elementary, middle, and high school). Most of the tests in this collection are in the subject areas of reading and mathematics, but there are neither strong nor consistent differences by subject so all subjects are combined in these summaries.

As Table 9 shows, the vast majority of available randomized studies examined interventions at the elementary school level, so the ability to make normative effect size comparisons for educational interventions in middle and high school is quite limited. There is, nonetheless, a relatively consistent pattern of smaller average effect sizes for interventions with high school students. The strongest pattern in Table 9 involves the type of achievement test. On average, larger effect sizes have been found across all grade levels on specialized researcher-developed tests, which are presumably more closely aligned with the intervention being evaluated, than on more general standardized tests. Moreover, the average effect sizes on the standardized tests are larger for more specific subtests than for the broadband overall achievement scores, with the latter being notably small. Based on the experience represented in these studies, rather modest effect sizes on broad achievement tests, such as the reading and mathematics tests required by state departments of education, could still be in the average or upper range of what any intervention has been able to produce on such measures.

Table 9. Achievement effect sizes from randomized studies broken out by type of test and grade level

Type of Test	Grade Level	N of Effect Sizes	Median	Mean	Standard Deviation
Specialized Topic or Test, Researcher Developed	Elementary	230	.34	.40	.55
	Middle	27	.35	.43	.48
	High	43	.29	.34	.38
	Total	300	.31	.39	.53
Standardized Test, Narrow Scope	Elementary	374	.17	.25	.42
	Middle	30	.26	.32	.26
	High	22	.04	.03	.07
	Total	426	.16	.24	.40
Standardized Test, Broad Scope	Elementary	89	.07	.08	.27
	Middle	13	.11	.15	.33
	High	1	--	--	--
	Total	103	.07	.08	.28
Total	Elementary	693	.19	.28	.46
	Middle	70	.21	.33	.38
	High	66	.09	.23	.34
	Total	829	.18	.28	.45

NOTE: Standardized mean difference effect sizes from 181 samples. No weighting was used in the calculation of the summary statistics and no adjustments were made for multiple effect sizes from the same sample.

Of course, we would expect that the nature of the intervention itself might be related to the magnitude of the achievement effects. Many different interventions are included in the collection of studies summarized here. To allow for some insight about the normative distributions of effect sizes for different kinds of

interventions, we categorized them in two ways. The first of these sorted the interventions according to their general nature into the following rather crude categories:

- *Instructional format*. Approaches or formats for instruction; broad pedagogical strategies having to do with how instruction is organized or delivered (e.g., cooperative learning, peer assisted learning, simulation games).
- *Teaching technique*. Specific pedagogical techniques or simple augmentations to instruction; not as broad as the approaches or strategies above; may be in one subject area but not specific to the content of a particular lesson (e.g., advance organizers before lectures, using hand calculators, using word processor for writing, providing feedback, mastery learning techniques, constructivist approaches to math lessons).
- *Instructional component or skill training*. Content-oriented instructional component, but of limited scope; distinct part of a larger instructional program, but not the whole program; training or instructional scheme focused on training a particular skill (e.g., homework, phonological awareness training, literature logs, repeated reading, computer-assisted tutoring on specific content; word analogy training).
- *Curriculum or broad instructional program*. A relatively complete and comprehensive package for instruction in a content area like a curriculum or a more or less freestanding program (e.g., science or math curriculum; reading programs for younger students; broad name brand programs like Reading Recovery; organized multisession tutoring program in a general subject area).
- *Whole school program*. Broad schoolwide initiatives aimed at improving achievement; programs that operate mainly at the school level, not via teachers' work in classrooms; whole school reform efforts (e.g., professional development for teachers, comprehensive school reform initiatives, Success for All).

Another categorization of the interventions was done that distinguished the nature of the group to which the educational intervention was targeted. This scheme used the following categories:

- *Individual students*. One-on-one tutoring, individualized instruction, and other interventions that worked with students individually.
- *Small group*. Interventions delivered to small groups of students, e.g., pull-out programs for poor readers.
- *Classroo*m. A curriculum or activity used by the teacher for the whole class such as cooperative learning or a math curriculum.
- *Whole school*. Interventions delivered to the whole school or with a schoolwide target such as comprehensive school reform initiatives or training for the teachers.
- *Mixed*. Interventions that involved activities targeted on more than one level; combinations of more than one of the above with none dominant.

Table 10 summarizes the data for the effect sizes in these categories found in the collection of randomized studies we are using to illustrate normative comparisons as a way to appraise the practical significance of effect sizes. The results there are not further broken out by type of achievement test and grade level (as shown in Table 9). Most of the studies, recall, were conducted with elementary students. Though there was some variation in the average effect sizes across test type and grade within both the type of intervention and target recipient categories, there were no strong or consistent differences from the overall aggregate results and many of the cells have too few effect sizes to support such close scrutiny.

Table 10 shows that, on average, larger achievement effect sizes have been found for interventions in the broad *teaching techniques* and *instructional component/skill training* categories than for other intervention approaches. Our point here is not to identify the kinds of interventions that are most effective, which would require a much more detailed analysis, but to provide a general picture of the order of magnitude of the effects that would be judged larger or smaller relative to such general norms. Similarly, Table 10 shows that the average effects for interventions that work with individual students, small groups, or involve activities that engage students in different groupings are generally larger than the average effects for interventions that target whole classrooms or whole schools. What would be large relative to such norms for interventions that target classrooms or schools might be only average or below for interventions that target individual students or small groups of students.

Table 10. Achievement effect sizes from randomized studies broken out by type of intervention and target recipients

	N of Effect Sizes	Median	Mean	Standard Deviation
Type of Intervention				
Instructional format	52	.13	.21	.36
Teaching technique	117	.27	.35	.47
Instructional component or skill training	401	.27	.36	.50
Curriculum or broad instructional program	227	.08	.13	.32
Whole school program	32	.17	.11	.31
Total	829	.18	.28	.45
Target Recipients				
Individual students	252	.29	.40	.53
Small group	322	.22	.26	.40
Classroom	176	.08	.18	.41
Whole school	35	.14	.10	.30
Mixed	44	.24	.30	.33
Total	829	.18	.28	.45

NOTE: Standardized mean difference effect sizes from 181 samples. No weighting was used in the calculation of the summary statistics and no adjustments were made for multiple effect sizes from the same sample.

Note that we are not suggesting that there is anything definitive about the results of the collection of randomized education studies presented here. Though these studies provide a broad picture of the magnitude and variability of mean effect sizes, their main purpose is to illustrate the approach of assessing the practical significance of an effect found in a particular intervention study by comparing it with the effects that have been found in broadly similar interventions with similar student samples. Normative effect size data that are more tailored for comparison with a particular intervention may be found by searching the research literature for the most similar studies and matching more carefully than can be done with the general summaries in Tables 9 and 10. Another potential source of effect size estimates appropriate for comparison with the findings from the intervention study at issue is the large body of meta-analysis research that has been done on educational interventions (e.g., see Hattie 2009). Those meta-analyses typically report not only mean effects sizes but also breakdowns by different outcomes, student subgroups, intervention variations, and so forth. When the standard deviations for mean effect sizes are also reported, the general shape of the full presumptively normal distribution can be reconstructed if desired so that, for instance, the percentile value of any given effect size can be estimated.

Benchmarking Effects Relative to Cost

Few approaches to assessing the practical significance of an intervention effect are more inherently meaningful to policy makers, citizens, and educational administrators than figuring the dollar cost of the intervention in relation to the effects that it produces. One instructive way to characterize intervention effects, therefore, is in terms of their cost-effectiveness or cost-benefit relationships. When used to describe intervention effects, these two forms of cost-analysis are distinguished from each other by the way the intervention effect is represented. For cost-effectiveness, the intervention effect is represented as an effect, that is, as an effect size or some analogous index of effect, such as the number of additional students scoring above the proficient level on an achievement test as a result of the intervention. When presented in cost-benefit terms, the intervention effect is converted into its expected dollar value. The dollar cost of the program can then be compared directly with the dollar value of the benefits in a cost-benefit ratio that indicates the expected financial return for a dollar spent.

Detailed guidance for conducting a cost analysis is beyond the scope of this paper, but below we outline the general approach and provide examples to illustrate how it can be used as a framework for assessing the practical significance of intervention effects. Fuller accounts can be found in Levin and McEwan (2001), Brent (2006), and Mishan and Quah (2007).

Calculating Total Cost

Cost-effectiveness and cost-benefit assessments of intervention effects begin by determining the cost of the intervention. The procedure typically employed for educational and other human service interventions is called the ingredients approach (Levin 1988; Levin and McEwan 2001). The ingredients approach involves three basic steps: identification of all the ingredients required to implement the intervention, estimation of the cost of those ingredients, and analysis of those costs in relation to the effects or monetary benefits of the intervention.

The ingredients necessary to implement an intervention include the personnel, materials, equipment, and facilities required as well as any associated indirect support such as maintenance, insurance, and the like. The ingredients identified for a classroom intervention, for instance, might include the paid and unpaid personnel who deliver the intervention (e.g., teachers, aides, specialists, and volunteers), the materials and equipment required (e.g., textbooks, curriculum guides, and software; computers, projectors, and lab equipment), the facilities that provide the setting (e.g., classroom, building, playground, and gymnasium), and the indirect support associated with these primary ingredients (e.g., janitorial services, electric and water utilities, and a share of the time of school administrative personnel). Everything required for sustaining the intervention is included, even those inputs that are not part of the intervention itself, like transportation to the intervention site (which might be provided by parents).

Once all of the necessary ingredients have been identified, the dollar value of each item must be estimated, even items like volunteer labor or donated materials that are not paid directly by the sponsors of the intervention. It is critical that all costs be systematically identified and this typically requires multiple data sources, e.g., written reports and records, interviews, and direct observations. There are different accounting frameworks for making the cost estimates, including actual cost as purchased, fair market value, and opportunity costs. And, there are various adjustments to be made to some of the costs, e.g., amortizing the cost of capital improvements, depreciation, discounting the value of future obligations, and adjusting for inflation (Levin and McEwan 2001). When all ingredients have been itemized and the cost of each separately estimated, their dollar values can be aggregated into a total cost estimate for the intervention or broken down into different categories of particular interest.

Table 11. Estimated costs of two fictional high school interventions

	After-school tutoring (50 students at 5:1 ratio for approximately one semester)	**Computer-assisted instruction** (Schoolwide computer lab with one instructor for 300 students)
Personnel	$75,000 10 certified teachers for 2hrs/day at $50/hr x 15 weeks	$50,500 Salary and benefits for full-time lab instructor
Facilities	$1,500 Utilities for 10 classrooms 2hrs/day	$2,000 Wiring and renovations for one classroom annualized for 5 years $1,250 Annual utilities for one classroom

Table 11. Estimated costs of two fictional high school interventions – continued

	After-school tutoring (50 students at 5:1 ratio for approximately one semester)	**Computer-assisted instruction** (Schoolwide computer lab with one instructor for 300 students)
Equipment & Supplies	$1,500 Photocopies and other incidentals	$6,000 30 computers annualized for 5 years $1,500 Laser printers and digital projector annualized for 3 years $7,500 Software annualized for 3 years $4,750 Annual equipment maintenance $1,500 Printer paper and other supplies
Professional development	$2,000 Teacher stipends for annual training	$6,000 Annual training workshops
Total Costs	$80,000	$81,000
Total Cost/ Student	$1,600	$270

Table 11 presents two fictional examples of educational interventions that illustrate the nature of an ingredients list and the associated itemized cost estimates. For each intervention, the major resources required are listed along with their estimated annual costs. Neither list exhaustively covers every necessary ingredient, as a full cost analysis would, but they suffice for demonstrating the approach. The costs of equipment and capital improvements expected to last longer than a year are annualized to spread those costs over their expected useful lifetime. Costs that recur annually are shown in total for a single year. For both intervention examples, the dollar amounts are assumed to be estimated by a credible procedure and adjusted if appropriate for depreciation, inflation, and the like. Although the computer-assisted instruction program has higher initial costs because of the need to purchase computers and software, this example shows that the total annualized cost could be comparable to that of the after-school tutoring intervention. But, because the after-school tutoring program serves fewer students, its cost per student is considerably higher than that for the computer-assisted instruction.

Cost-effectiveness

Once total costs have been calculated, cost effectiveness estimates can be obtained for any number of different representations of the effects of an intervention. As discussed throughout this paper, there are many ways that such effects might be represented, each of which is associated with a different unit of measurement. For instance, the unit of effect could be in the original metric, e.g., the mean number of points per student gained on the annual state mandated mathematics achievement test. If that mean gain was, say, 50 points for the after-school tutoring program in Table 11, we could construct a cost effectiveness ratio for the gain expected for some standard cost per student, e.g., gain for each $1,000 spent per student (Harris 2009). With the cost to deliver the after-school tutoring program estimated at $1,600 per student, the cost effectiveness ratio for that program would be about 31, that is, a mean gain of 31 points on the test per $1,000 spent on a student participating in the program ([1,000/1,600] x 50). Characterizing the intervention effect in terms of what is gained for a given per student dollar investment can make its practical significance easier to assess in the context of application.

This way of representing intervention effects will be especially meaningful for outcomes such as graduation rates that are widely-understood and provide intuitive metrics for policy-makers, administrators, and the general public. The practical significance of the increase in the percentage of students who graduate for a $1,000 per student investment in a program is readily understood by almost everyone. No matter what the overall effect is, if it works out to, say, one-tenth a percentage point increase with a per student cost of $1,000, few stakeholders will find it impressive. Similarly, some tests (e.g., the SAT college admissions test) are familiar to broad audiences so that the number of points on average by which a student's score is raised for a given cost in dollars may be easily appreciated.

In many instances, the cost-effectiveness results may be better understood when the intervention effects are represented in terms of standardized effect sizes, especially when those results are available in a comparative framework. For example, cost-effectiveness ratios with effect sizes as the unit in which the intervention effect is represented can be illustrated for the fictional school tutoring and computer-assisted instruction interventions in Table 11. Suppose that the effects on the scores of 10th grade students on the state administered mathematics achievement test and on their graduation rates are estimated in a number of high schools. The effect on the math scores can be represented in standard deviation units as a standardized mean difference effect size with the after-school tutoring producing an effect size of, say, .35, and the computer-assisted instruction producing an effect size of .20. Graduation rates are typically reported as the percent of students who graduated, making this a binary variable for which an appropriate effect size statistic is the odds ratio. For instance, if 80% of the students in the control group graduate compared to 83% in the intervention group, the intervention effect expressed as an odds ratio is 1.22—the odds of an intervention student graduating (83/17) are 22% greater than the odds of a control student graduating (80/20) (see section 3.1.1). Suppose then that the odds ratio for the effect on graduation rates for the after-school program is 1.22 and that for the computer-assisted instruction is 1.10.

As Table 12 shows, the after-school tutoring program yields an increase of .22 standard deviations in the math achievement test score (that is, an effect size of .22) for a $1,000 per student cost. The comparable cost-effectiveness ratio for the computer-assisted instruction is .74. Looked at individually, these effect size

values in standard deviation units will have limited meaning to most audiences, but viewed comparatively it is easy to see that the computer-assisted instruction program provides more bang for the buck. Similarly for graduation rates, the after-school program produces a 14 percentage point increase in the odds of graduation for a per student investment of $1,000 whereas the computer-assisted instruction program produces a 37 point increase for the same investment. These examples highlight the most notable advantage of comparative cost-effectiveness analysis, which is that it allows consumers to see that the most effective intervention may not be the most cost-effective one.

Table 12. Cost-effectiveness estimates for two fictional high school interventions

	10th Grade Math Scores		Graduation Rates	
	After-school tutoring	**Computer-assisted instruction**	**After-school tutoring**	**Computer-assisted instruction**
Intervention effect	.35 ES	.20 ES	1.22 Odds-ratio 22% increase in odds	1.10 Odds-ratio 10% increase in odds
Cost/student	$1,600	$270	$1,600	$270
Cost-effectiveness ratio	.22 SD of effect for $1,000 per student	.74 SD of effect for $1,000 per student	14 % increase in odds for $1,000 per student	37% increase in odds for $1,000 per student

In Section 4.3 above, we described how the practical significance of the effect size found in a given intervention study could be assessed by comparing it with the effect sizes reported in other studies of similar interventions with similar populations and outcomes. Harris (2009) suggested going a step further in such comparative assessments. If the cost per student of the respective interventions were also routinely reported, cost-effectiveness ratios such as those in Table 12 would be available for different outcomes across a variety of interventions. The practical significance of an effect found in an intervention study could then also be assessed by comparing its cost-effectiveness ratio with the cost-effectiveness ratios found in other studies on similar outcomes. All else equal, the interventions that produce the largest effects for a given per student cost have the greatest practical significance in a world of limited resources.

Additional Considerations and Limitations of Cost-effectiveness Estimates

Many variants are possible when estimating the cost of a program in cost effectiveness analysis. One is to partition the costs according to who is responsible for paying them. Each contributing party may be most interested in the costs it supports. For instance, a school district may exclude costs that are paid by federal and state funding sources from its estimates and include only the costs to the district. Similarly, it may be of interest to estimate the costs separately for different sites or different subgroups participating in the intervention when those costs are expected to vary. If intervention effects are smaller for some groups, the cost effectiveness of the intervention will be less for them unless the costs per student are also decreased proportionately. Cost per student, of course, may also vary if different groups of students require different

support for the intervention. Students in special education, for instance, may have different costs per hour of instruction for a given type of instruction because of the need for higher teacher to student ratios in the classroom.

Similarly, rather than total costs, cost estimates might be generated for marginal costs—the additional cost beyond current business-as-usual when a new intervention is implemented (Levin and McEwan 2001). For instance, much of the cost associated with a new teaching method may be the same as for current practice with only additional training required for successful implementation. Decision makers may want to assess the effect of the new method relative to the old in terms of the improvement in outcome attained for the additional cost of the new one. In a study of the effects of the Success for All (SFA) program, for instance, Borman and Hewes (2002) estimated the per student cost for both the SFA intervention in the intervention schools and practice-as-usual in the control schools. Those costs were not significantly different, indicating that the positive outcomes of SFA were attained for no net additional cost to the school district.

Although cost-effectiveness estimates can provide a relatively direct and intuitively meaningful way to assess the practical magnitude of an intervention effect, this approach does have some inherent limitations. The quality of the research that estimates the intervention effects directly affects the precision and reliability of cost-effectiveness estimates and poor effect estimates will yield poor cost effectiveness estimates. Moreover, deciding which costs are relevant and how to estimate them involves a series of judgment calls that different analysts may make differently. Decisions on those matters can significantly alter the cost-effectiveness estimates that result. Levin (1988) suggested that cost-effectiveness analysis is so imprecise that estimates should differ by at least 20% to be meaningful. However, this guideline itself is imprecise and a number of professional boards in the medical, health, and education fields have recommended that cost-effectiveness analyses always be complemented by additional evaluations, metrics, and sensitivity analysis (Hummel-Rossi and Ashdown 2002).

Cost-benefit

Like cost-effectiveness, cost-benefit assessments begin with estimation of the total cost of the intervention, but then estimate the economic value of the effect or effects produced by the intervention as well. An effect such as an increased graduation rate as a result of a dropout prevention program is thus converted to a dollar value, known as a shadow price or dollar valuation (Karoly 2008). As with the estimation of intervention costs, estimates of the value of the benefits must be represented as real monetary values by adjusting for inflation and converting future benefits to current dollar values. Once total benefits are calculated, the ratio of benefits to costs can be calculated to yield the dollar value of the intervention effects per dollar spent to obtain that benefit. The practical significance of the intervention effect can then be appraised in terms of that ratio. If a dollar spent on the intervention returns more than a dollar's worth of benefit, especially if it returns considerably more than that, the effects of that intervention can be viewed as large enough to have practical value.

The process of identifying and costing the benefits of an intervention involves a number of decisions about which implications of the effects to include in estimating the benefits. Additionally, many of the benefits of educational interventions are not typically valued in dollars and it can be challenging to develop a

plausible basis for such estimates. Consequently, not all potential benefits of an educational intervention lend themselves readily to cost-benefit analysis. For instance, Levin, Belfield, Muennig, and Rouse (2007) calculated the costs and benefits of five interventions designed to increase high school graduation rates. Levin et al. chose to include as benefits to the taxpayers who support public education such consequences of graduation as contributions to tax revenues, reduction in public health costs, and reductions in criminal justice and public welfare costs. However, they did not include other possible benefits, such as an informed citizenry, because there was too little basis in available data to assign a dollar value to that effect. It is typical of cost-benefit estimates for education intervention effects that the outcomes are not directly monetized but, rather, are used to infer indirect monetary impacts that follow from those effects. Improvements in achievement test scores, for instance, are commonly monetized on the basis of the relationship between achievement and future earnings (Karoly 2008).

Cost-benefit analysis also requires consideration of how costs and benefits are distributed among stakeholders. For instance, cost-benefit estimates can be made separately for personal benefits and societal benefits. Lifetime earnings and additional educational attainment are common examples of personal benefits, while reduction in crime and health care costs are typically considered public benefits. It is often the case that the benefits do not accrue to different stakeholders in proportion to their costs. Thus an educational intervention can have a very favorable overall benefit-cost ratio but an unfavorable one when only the costs and benefits to, e.g., the local state budget are considered. Also, benefits may be different for different sub-groups of participants. For example, Levin et al. (2007) calculated cost and benefits separately both for personal and public benefits and for racial and gender subgroups receiving interventions designed to increase high school graduation rates.

Cost-benefit analysis is a useful tool for decision makers, but proponents caution that it should not be the sole basis for assessing the effects of educational interventions. Considerations other than cost often apply to the social value placed on educational outcomes. A related caution is primarily methodological. There are consequential judgment calls that must be made for analysts to convert intervention effects into dollar amounts and different researchers can use different methods and produce different results for the same intervention. Cost-benefit estimates, therefore, can vary widely across analysts and studies. Moreover, evidence from other fields suggests that costs are frequently underestimated and benefits are frequently overestimated. These errors can result in benefit-cost ratios that are positively biased and thus overstate the practical significance of the respective intervention effects (Flyvbjerg, Holm, and Buhl 2002, 2005).

References

Albanese, M. (2000). Problem-Based Learning: Why Curricula are Likely to Show Little Effect on Knowledge and Clinical Skills. *Medical Education*, *34*: 729-738.

Belfield, C. R., and Levin, H. M. (2009). *High School Dropouts and the Economic Losses from Juvenile Crime in California*. Santa Barbara, CA: University of California, Santa Barbara.

Bloom, H. S., Hill, C. J., Black, A. B., and Lipsey, M. W. (2008). Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions. *Journal of Research on Educational Effectiveness*, *1*(4): 289-328.

Borman, G. D., and Hewes, G. M. (2002). The Long-Term Effects and Cost-Effectiveness of Success for All. *Educational Evaluation and Policy Analysis*, *24*(4): 243-266.

Brent, R. J. (2006). *Applied Cost-Benefit Analysis* (2nd ed.). Northhampton, MA: Edward Elgar Publishing.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (revised edition). New York: Academic Press.

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement*, *7*(3): 249-253.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crichton, N. (2001). Information Point: Odds Ratios. *Journal of Clinical Nursing*, *10*(2): 268–269.

CTB/McGraw Hill. (1996). *California Achievement Tests*, Fifth Edition. Monterey, CA: Author.

Decker, P. T., Mayer, D. P., and Glazerman, S. (2004). The Effects of *Teach for America on Students: Findings from a National Evaluation*. Princeton, NJ: Mathematica Policy Research, Inc. MPR Reference No. 8792-750.

Dunn, L.M., and Dunn, L.M. (2007). *Peabody Picture Vocabulary Test-Fourth Edition*. Bloomington, MN: Pearson Assessments.

Flyvbjerg, B., Holm, M. K. S., and Buhl, S. L. (2002). Underestimating Costs in Public Works Projects. Error or Lie? *Journal of the American Planning Association*, *68*(3): 279-295.

Flyvbjerg, B., Holm, M. K. S., and Buhl, S. L. (2005). How (In)accurate are Demand Forecasts in Public Works Projects? The Case of Transportation. *Journal of the American Planning Association*, *71*(2): 131-146.

Harcourt Assessment, Inc. (2004). *Stanford Achievement Test Series*, *Tenth Edition Technical Data Report*. San Antonio, TX: Author.

Harcourt Educational Measurement. (1996). *Stanford Achievement Test Series*, *9th Edition*. San Antonio, TX: Author.

Harris, D. N. (2009). Toward Policy-Relevant Benchmarks for Interpreting Effect Sizes: Combining Effects with Costs. *Educational Evaluation and Policy Analysis*, *31*(1): 3-29.

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. NY: Routledge.

Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics*, *32*(4): 341-370.

Hedges, L. V., and Hedberg, E. C. (2007). Intraclass Correlation Values for Planning Group Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, *29*(1): 60-87.

Hills, J. R. (1984). Interpreting NCE Scores. *Educational Measurement: Issues and Practice*, *3*(3): 25-26, 31.

Hummel-Rossi, B., and Ashdown, J. (2002). The State of Cost-Benefit and Cost-Effectiveness Analyses in Education. *Review of Educational Research*, *72*(1): 1-30.

Karoly, L. A. (2008). *Valuing Benefits in Benefit-Cost Studies of Social Programs*. Santa Monica, CA: RAND.

Konstantopoulos, S., and Hedges, L. V. (2008). How Large an Effect can We Expect from School Reforms? *Teachers College Record*, *110*(8): 1613–1640.

Levin, H. M. (1988). Cost-Effectiveness and Educational Policy. *Educational Evaluation and Policy Analysis*, *10*(1): 51-69.

Levin, H. M. (1995). Cost-Effectiveness Analysis. In M. Carnoy (ed.), *International Encyclopedia of Economics of Education* (2nd ed.; pp. 381-386). Oxford: Pergamon.

Levin, H. M., Belfield, C., Muennig, P., and Rouse, C. (2007). The Public Returns to Public Educational Investments in African American Males. *Economics of Education Review*, *26*(6): 699-708.

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

Linn, R. L., and Slinde, J. A. (1977). The Determination of the Significance of Change Between Pre- and Post-Testing Periods. *Review of Educational Research*, *47*(1): 121-150.

Lipsey, M. W., and Wilson, D. B (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications.

MacCallum, R.C., Zhang, S., Preacher, K.J., and Rucker, D.D. (2002). On the Practice of Dichotomization of Quantitative Variables. *Psychological Methods*, *7*(1): 19-40.

McCartney, K., and Rosenthal, R. (2000). Effect Size, Practical Importance, and Social Policy for Children. *Child Development*, *71*(1): 173-180.

Mishan, E. J., and Quah, E. (2007). *Cost-Benefit Analysis* (5th edition). New York, NY: Routledge.

Ramos, C. (1996). The Computation, Interpretation, and Limits of Grade Equivalent Scores. Paper presented at the Annual Meeting of the Southwest Educational Research Association, New Orleans, LA.

Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic Methods for Questions Pertaining to a Randomized Pretest, Posttest, Follow-up Design. *Journal of Clinical Child & Adolescent Psychology*, *32*(3): 467-486.

Redfield, D. L., and Rousseau, E. W. (1981). A Meta-Analysis of Experimental Research on Teacher Questioning Behavior. *Review of Educational Research*, *51*(2): 237-245.

Roderick, M., and Nagaoka, J. (2005). Retention Under Chicago's High-Stakes Testing Program: Helpful, Harmful, or Harmless? *Educational Evaluation and Policy Analysis*, *27*(4): 309–40.

Rosenthal, R., and Rubin, D. B. (1982). A Simple, General Purpose Display of Magnitude of Experimental Effect. *Journal of Educational Psychology*, *74*(2): 166-169.

Slavin, R. E., Madden, N. A., Lawrence, J. D., Wasik, B. A., Ross, S., Smith, L., and Dianda, M. (1996). Success for All: A Summary of Research. *Journal of Education for Students Placed at Risk*, *1*(1): 41-76.

Tallmadge, G.K., and Wood, C.T. (1976). *ESEA Title I Evaluation and Reporting System: User's Guide*. Mountain View, CA: RMC Research Corp. ERIC ED 309195.

What Works Clearinghouse. (2011). *Procedures and Standards Handbook* (version 2.1). U.S. Department of Education. Washington, DC: Institute of Education Sciences. Retrieved October 12, 2012 from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf.