Empirical Benchmarks for Interpreting Effect Sizes in Research

Carolyn J. Hill,[1] Howard S. Bloom,[2] Alison Rebeck Black,[2] and Mark W. Lipsey[3]

[1]Georgetown Public Policy Institute, [2]MDRC, and [3]Vanderbilt Institute for Public Policy Studies

ABSTRACT—***There is no universal guideline or rule of thumb for judging the practical importance or substantive significance of a standardized effect size estimate for an intervention. Instead, one must develop empirical benchmarks of comparison that reflect the nature of the intervention being evaluated, its target population, and the outcome measure or measures being used. This approach is applied to the assessment of effect size measures for educational interventions designed to improve student academic achievement. Three types of empirical benchmarks are illustrated: (a) normative expectations for growth over time in student achievement, (b) policy-relevant gaps in student achievement by demographic group or school performance, and (c) effect size results from past research for similar interventions and target populations. The findings can be used to help assess educational interventions, and the process of doing so can provide guidelines for how to develop and use such benchmarks in other fields.***

KEYWORDS—*effect size; student performance; educational evaluation*

Studies of treatment effectiveness abound across a broad range of program areas. In education, for example, studies have examined whether preschool interventions increase school readiness (e.g., Magnuson, Ruhm, & Waldfogel, 2007), whether curricular interventions increase reading or mathematics achievement (e.g., Snipes, Holton, & Doolittle, 2006), and whether after-school programs reduce dropout from high school (e.g., Dynarski, Gleason, Rangarajan, & Wood, 1998). Tests of statistical significance for estimated treatment effects in these studies provide insight into whether the observed effects might have occurred by chance alone. Yet these tests do not provide insight into whether the *magnitudes* of effects are substantively or practically important—an issue of particular interest to policy makers and program officials.

Translating the estimated effect of an intervention into a standardized *effect size*—calculated as the difference in means between treatment and control groups, divided by the pooled standard deviation of the two groups—provides one way to interpret the substantive significance of interventions. Standardized effect size measures, unlike tests of statistical significance, are independent of sample size. Typically, these effect size magnitudes have been interpreted on the basis of rules of thumb suggested by Cohen (1988), where an effect size of approximately 0.20 is considered "small," approximately 0.50 is considered "medium," and approximately 0.80 is considered "large." Although the Cohen guidelines are broad generalizations covering many types of interventions, target populations, and outcome measures, it has become standard practice for researchers and policy makers to use them when interpreting effect size estimates.

We argue that effect sizes should instead be interpreted with respect to empirical benchmarks that are relevant to the intervention, target population, and outcome measure being considered. We have presented the data and points we describe here at numerous seminars and conferences and in working papers (Bloom, Hill, Black, & Lipsey, 2006a, 2006b, 2007a, 2007b, in press; Lipsey, Bloom, Hill, & Black, 2007). We use the data to present statistical points, but because all the findings from the full research project have not yet undergone formal peer review, the substantive content of information shown in the tables should be considered preliminary.

The research on which this article is based received support from the Institute of Education Sciences in the U.S. Department of Education, the Judith Gueron Fund at MDRC, and the William T. Grant Foundation. The authors thank Larry Hedges for his helpful input.

Correspondence concerning this article should be addressed to Howard S. Bloom, MDRC, 16 East 34th Street, 19th Floor, New York, NY 10016-4326; e-mail: howard.bloom2@mdrc.org.

We illustrate our points with three of the numerous possible types of benchmarks: (a) normative expectations for change, (b) policy-relevant performance gaps, and (c) effect size results from similar studies. Our analysis draws from a larger ongoing research project that is examining the calculation, interpretation, and uses of effect size measures in education research. Thus, we illustrate each benchmark with educational examples. The more general message—that effect sizes should be interpreted using relevant empirical benchmarks—is applicable to any policy or program area, however.

BENCHMARK 1: NORMATIVE EXPECTATIONS FOR CHANGE

Our first empirical benchmark refers to expectations for growth or change *in the absence of an intervention*. In the context of education, the question is: How does the effect of an intervention compare to a typical year of growth for a given target population of students?

To explore this issue, we build on an approach developed by Kane (2004). Our analysis uses test scores from kindergarten to 12th grade for the national norming samples of seven major standardized tests in reading and six tests in math (Bloom et al., 2006a, 2006b, 2007a, 2007b, in press; Lipsey et al., 2007).[1] These tests are similar to other broadband achievement tests widely used at the state and national levels. We used each test's technical manuals to obtain its mean scale score and student-level standard deviation, by grade.[2] (These scores are designed for comparisons across grades.) For each test, we measure annual growth in achievement by calculating the difference of mean scale scores in adjacent grades. We then convert the difference to a standardized effect size by dividing it by the pooled student-level standard deviation for the two adjacent grades. Finally, we aggregate information across tests by taking the random-effect mean effect size for each grade-to-grade transition, with weights based on the formulas in Hedges (1982). These estimates are measured from spring to spring and thus represent learning gains from a "year of life," which captures learning in school, learning and maturation outside of school, plus any learning loss experienced during the summer.[3]

The resulting growth trajectories for reading and math effect sizes are shown in the "Mean" columns of Table 1. The margin of error (for a 95% confidence interval [CI]) for each estimate is shown in parentheses. For example, the average annual reading gain measured in effect size from Grade 1 to Grade 2 is 0.97 *SD*. Because the margin of error for this estimate is 0.10, the lower bound of its 95% CI is 0.87 and the upper bound is 1.07.

The trajectories of annual gains in Table 1 exhibit a strikingly consistent pattern for both reading and math. Gains are largest in the lower elementary grades and then decline steadily into the high school years. For example, the average annual reading gain is 0.97 *SD* for Grades 1–2, 0.32 *SD* for Grades 5–6, and only 0.06 *SD* for Grades 11–12. Although the estimates do not always decrease from year to year, the overall trend is clear: The natural growth in test scores declines as students age. We observed the same pattern of findings for tests of social studies and science (Bloom et al., in press). Because the standard deviations are relatively stable across grades, the pattern is driven primarily by a decreasing difference in means.

Before interpreting the findings in Table 1, it is important to consider some caveats about them. First, these findings may partly reflect an inconsistency between the material being taught and the material being tested in upper grades. Second, the sample composition for upper grades is changing across grades because of students who drop out of school. Third, the patterns in Table 1 for national norming samples may differ from those for local school districts or subgroups of students. Fourth, the mean differences between grades may differ for

Table 1

Average Annual Gain in Effect Size From Nationally Normed Tests

	Reading tests		Math tests	
Grade transition	Mean	Margin of error	Mean	Margin of error
Grade K–1	1.52	±0.21	1.14	±0.49
Grade 1–2	0.97	±0.10	1.03	±0.14
Grade 2–3	0.60	±0.10	0.89	±0.16
Grade 3–4	0.36	±0.12	0.52	±0.14
Grade 4–5	0.40	±0.06	0.56	±0.11
Grade 5–6	0.32	±0.11	0.41	±0.08
Grade 6–7	0.23	±0.11	0.30	±0.06
Grade 7–8	0.26	±0.03	0.32	±0.05
Grade 8–9	0.24	±0.10	0.22	±0.10
Grade 9–10	0.19	±0.08	0.25	±0.07
Grade 10–11	0.19	±0.17	0.14	±0.16
Grade 11–12	0.06	±0.11	0.01	±0.14

Sources. Annual gain for reading is calculated from seven nationally normed tests: California Achievement Test (CAT)–5th edition, Stanford Achievement Test (SAT)–9th edition, TerraNova–Comprehensive Test of Basic Skills (CTBS), Metropolitan Achievement Test (MAT8), TerraNova–CAT, SAT10, and Gates-MacGinitie. Annual gain for math is calculated from six nationally normed tests: CAT5, SAT9, TerraNova–CTBS, MAT8, Terra Nova–CAT, and SAT10. For further details, contact the authors (Bloom et al. 2006a, 2006b, 2007a, 2007b, in press; Lipsey et al., 2007).

[1] California Achievement Test (CAT)–5th edition, Stanford Achievement Test Series (SAT)–9th edition, TerraNova–Comprehensive Test of Basic Skills, Gates-MacGinitie, Metropolitan Achievement Test (MAT8), TerraNova–CAT, and SAT Series–10th edition. The Gates-MacGinitie does not include a mathematics component.

[2] The decision of whether to use the student-level or school-level standard deviation to calculate the effect size is a consequential one, but beyond the scope of this article.

[3] These are cross-sectional estimates. However, using data for individual students from several large urban school districts, we find that gains calculated from longitudinal data (year-to-year changes for the same students) are very similar to those calculated from cross-sectional data (grade-to-grade differences for a given year), except for the transition from 9th to 10th grade, when large numbers of students drop out of school (Bloom et al., in press).

different student subgroups, some of which may lag behind these norms.

Nevertheless, because the preceding pattern findings are so striking and consistent, it is reasonable to use them as benchmarks for interpreting effect size estimates from intervention studies. For example:

- A particular effect size from an intervention study—such as an effect size of 0.10 *SD*—would constitute a relatively smaller substantive change for students in early grades than for students in later grades. Thus, for some purposes, it may be important to interpret a study's effect size estimate in the context of natural growth for its target population. This point does *not* imply that it is necessarily *easier* to produce a given effect size change—say, of 0.10—for early grades than for later grades.
- Reading and math effect sizes for the nationally normed tests are sometimes similar and are sometimes different for a given grade, even though their overall trajectories are very similar. Thus, it is important to interpret a study's effect size estimate in the context of the outcome being measured.

BENCHMARK 2: POLICY-RELEVANT PERFORMANCE GAPS

Our second proposed type of empirical benchmark refers to *policy-relevant performance gaps*. In the context of education, the question here is: How do the effects of an intervention compare with existing differences among subgroups of students or schools? Konstantopoulos and Hedges (2008) illustrate such comparisons using data for nationally representative samples. Here, we describe the reasoning behind this procedure and present some examples.

Because often "the goal of school reform is to reduce, or better, eliminate, the achievement gaps between minority groups such as Blacks or Hispanics and Whites, rich and poor, and males and females . . . it is natural then, to evaluate reform effects by comparing them to the size of the gaps they are intended to ameliorate" (Konstantopoulos & Hedges, 2008, p. 1615). We illustrate such gaps in Table 2, which shows differences in reading and math performance for subgroups of a nationally representative sample using published findings from the National Assessment of Educational Progress (NAEP; Bloom et al., 2006a, 2007a, 2007b, in press; Lipsey et al., 2007). Gaps in reading and math scores are presented by race/ethnicity, income (free or reduced-price lunch status), and gender for the most recent NAEP assessments in Grades 4, 8, and 12. These gaps are measured in terms of effect sizes, that is, the difference in mean scores divided by the standard deviation of scores for all students in a grade.

For example, Black fourth graders scored 0.83 *SD* lower than White fourth graders on the reading assessment and 0.99 *SD* lower on the math assessment. A gap between Blacks and Whites is observed at each of the three grade levels, though it is somewhat smaller in 12th grade. The gaps between Hispanic and White students, and between students who were and were not eligible for a free or reduced-price lunch, show similar trends but are typically smaller than the corresponding Black–White gap. Finally, male students tend to score lower than females in reading but higher in math. These gender gaps are typically much smaller than corresponding race/ethnicity or income gaps.

The preceding gaps for a nationally representative sample of students may differ from their counterparts for a particular state or school district. Furthermore, gaps for a different outcome measure (e.g., a state-developed test) may differ from those presented. Nevertheless, the findings in Table 2 illustrate the following points about empirical benchmarks for effect sizes based on policy-relevant gaps:

- A particular effect size from an intervention study—for example, an effect size of 0.10—may constitute a smaller substantive change relative to some academic gaps (e.g., that for Blacks and Whites) than to others (e.g., that for males and females). Thus, for some purposes, it may be important to

Table 2
Demographic Performance Gap in Mean NAEP Scores, by Grade (in Effect Size)

Subject and grade	Black–White	Hispanic–White	Eligible–ineligible for free/reduced-price lunch	Male–Female
Reading				
Grade 4	−0.83	−0.77	−0.74	−0.18
Grade 8	−0.80	−0.76	−0.66	−0.28
Grade 12	−0.67	−0.53	−0.45	−0.44
Math				
Grade 4	−0.99	−0.85	−0.85	0.08
Grade 8	−1.04	−0.82	−0.80	0.04
Grade 12	−0.94	−0.68	−0.72	0.09

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 (Bloom et al., 2007a, 2007b, in press; Lipsey et al., 2007).

interpret a study's effect size estimate in the context of the subgroups of interest. This point does *not* imply that it is necessarily *easier* to produce a given effect size change—say, of 0.10—to close the gaps for some groups than for others.

- Policy-relevant gaps for demographic subgroups (e.g., the Black–White gap) may differ for different types of outcomes (here, reading and math) and for different stages of development (here, Grades 4, 8, and 12). Thus, it is important to interpret a study's effect size estimate in relation to the policy-relevant gap for a particular outcome measure and target population.
- Performance gaps can provide a relevant benchmark for effect sizes from interventions even if those interventions are not explicitly intended to reduce a particular performance gap.

In addition to policy-relevant gaps between students, performance differences between schools may also be relevant for policy. As Konstantopoulos and Hedges (2008) put it, because some "school reforms are intended to make all schools perform as well as the best schools . . . it is natural to evaluate reform effects by comparing them to the differences (gaps) in the achievement among schools in America" (p. 1615).

Table 3 illustrates these kinds of gaps (Bloom et al., 2006a, 2007a, 2007b, in press; Lipsey et al., 2007). However, instead of comparing low-performing schools to high-performing schools, it illustrates a gap that might be closed more easily: that between low-performing schools and "average" schools. To compute these gaps, we used individual student-level data from four large urban school districts. Between-school gaps are shown in reading and math test scores for Grades 3, 5, 7, and 10. For each grade in each school, we calculate the adjusted mean performance (using nationally normed standardized tests) over a 2- or 3-year period.[4] The means are regression adjusted for test scores in the prior grade and students' demographic characteristics (race/ethnicity, gender, age, and free-lunch status). We then generate the distribution of average school performance for each grade in each district. The cell values in Table 3 are the differences—measured in terms of an effect size based on student-level standard deviations—between low-performing schools (i.e., those at the 10th percentile for the specified grade in the district) and average-performing schools (i.e., those at the 50th percentile for a specified grade in the district).[5]

Table 3 illustrates, for example, that the reading test score gap (controlling for prior performance and demographic characteristics) between low- and average-performing schools in Grade 3 for District I is about 0.31 *SD*. The gap is larger in Grade 5 and smaller in Grades 7 and 10. The magnitudes and patterns of gaps for math in District I are similar to those for reading. The other districts included in the table exhibit similar patterns, although specific gaps vary across districts.

Table 3

Performance Gap in Effect Size Between "Average" and "Weak" Schools (50th and 10th percentiles)

	District findings			
Subject and grade	I	II	III	IV
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

Note. "NA" indicates that a value could not be computed because of missing test score data. Means are regression adjusted for test scores in prior grade and students' demographic characteristics.
Sources. Iowa Tests of Basic Skills (ITBS) for District I, Stanford Achievement Test (SAT)9 for District II, Metropolitan Achievement Test for District III, and SAT8 for District IV. See description in text for further details on the sample and calculations (Bloom et al., 2006a, 2007a, 2007b, in press; Lipsey et al., 2007).

Findings in Table 3 illustrate the following:

- A particular effect size (e.g., 0.10 *SD*) may be relatively small or large depending on the empirical benchmark that is most relevant. For example, such an effect size would be relatively large for Grade 3 in District III but relatively smaller for Grade 5 or 7.
- Effect sizes from particular studies might be usefully interpreted by comparing them with an empirical benchmark that relates "weak" to "average" (or "high") performance of organizations or institutions. In education research, such a benchmark is particularly relevant for whole-school reforms or grade-specific interventions.
- Benchmarks derived from local sources (such as school district data) may provide a relevant guide for interpreting effect sizes from particular studies instead of, or in addition to, findings from national-level data.

BENCHMARK 3: OBSERVED EFFECT SIZES FOR SIMILAR INTERVENTIONS

Our third empirical benchmark refers to effects observed previously for *similar types of interventions*. In the context of education research, the question is, How do the effects of an intervention compare to those from previous studies for similar grade levels, interventions, and outcomes? This approach uses

[4]We use the following standardized tests: for District I, scale scores from the Iowa Tests of Basic Skills (ITBS); for District II, scale scores from the SAT9; for District III, normal curve equivalent scores from the MAT; and for District IV, normal curve equivalent scores from the SAT8.

[5]The effect size of the difference between "average" and "weak" schools (at the 50th and 10th percentiles) in a district is calculated as 1.285 times the square root of the regression-adjusted school-level variance ()($\hat{\tau}^2$), divided by the unadjusted student-level standard deviation ($\hat{\sigma}$). Thus, gaps are computed for inferred points in the school performance distribution.

results from research synthesis or meta-analysis. We illustrate it with two such analyses.

The first summarizes estimates of achievement effect sizes from random assignment studies of educational interventions (Bloom et al., 2007b). These results are thus based on the most rigorous impact design available (Shadish, Cook, & Campbell, 2002). We identified 61 random assignment studies (reporting on 95 independent subject samples) published since 1995 that examined the effects of educational interventions on mainstream students. The synthesis does not include studies of special education students or of students with clinical problems, nor does it include studies of interventions targeted primarily toward behavioral problems. Furthermore, control groups had to have experienced "treatment as usual" (not an alternative treatment), and attrition of the sample had to be less than 20%. Because most studies report multiple effect size estimates (e.g., for multiple outcomes or grades), a total of 468 effect sizes (calculated using the student-level standard deviation) were summarized.

Table 4 presents these findings by grade level (elementary, middle, and high school). Findings for elementary school are also subdivided by type of outcome measure: standardized tests that cover a broad subject matter (such as the SAT9 composite reading test), standardized tests that focus on a narrow topic (such as the SAT9 vocabulary test), or 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).

Most of the available randomized studies examined interventions at the elementary school level. The mean effect size for these interventions is 0.33 *SD*; the corresponding mean effect size for middle schools is 0.51 *SD*, and that for high schools is 0.27 *SD*. Within studies of elementary schools, mean effect sizes are highest for specialized tests (0.44), next highest for narrowly focused standardized tests (0.23), and lowest for broadly focused standardized tests (0.07). These findings raise important issues about how the test used to measure the effectiveness of an educational intervention might influence the results obtained. However, some of the differences in findings may also be due to differences in interventions and target populations represented in each of the three outcome groups.

Our second example of an empirical benchmark from a research synthesis is a "meta-analysis of meta-analyses" (Bloom et al., 2007b; Lipsey et al., 2007). These findings summarize the results of 76 meta-analyses of past studies of educational interventions in kindergarten through 12th grade that reported mean achievement effect sizes (calculated using student-level standard deviations) for experimental and quasi-experimental studies and that provided some breakdown by grade range (elementary, middle, and high school). We located a total of 192 meta-analyses of educational interventions. These 76 are the subset that does not involve duplicate coverage of studies, provides a breakdown by grade range, and includes comparison group studies only (no before or after studies or correlational studies). When more than one meta-analysis provided mean effect size estimates for a given type of intervention, we computed a weighted mean (weighting by the number of studies included in each meta-analysis).

Table 5 reports descriptive statistics from this meta-analysis of meta-analyses. Averaged over the many different interventions, studies, and achievement outcomes encompassed in these meta-analyses, the mean effect sizes are in the 0.20–0.30 range. Moreover, there is remarkably little variation in the means across grade levels, despite considerable variation in the interventions and outcomes represented for the different grades.

Tables 4 and 5 illustrate the following points with regard to assessing the magnitudes of effect sizes from particular studies based on findings from related research:

- Empirical benchmarks from a research synthesis do not indicate what effects are *desirable* from a policy standpoint. Instead, they provide a snapshot of effects found in previous studies, that is, what might be *attainable*.
- Different ways of measuring the same outcome construct—for example, different kinds of achievement tests—may result in

Table 4
Summary of Effect Sizes From Randomized Studies

Achievement measure	Number of effect size estimates	Mean effect size	*SD*
Elementary school	389	0.33	0.48
Standardized test (broad)	21	0.07	0.32
Standardized test (narrow)	181	0.23	0.35
Specialized topic/test	180	0.44	0.49
Middle school	36	0.51	0.49
High school	43	0.27	0.33

Note. Unweighted means across all effect sizes and samples in each category.
Sources. Compiled by the authors from 61 existing research reports and publications (reporting on 95 independent subject samples; Bloom et al., 2007b).

Table 5
Distributions of Mean Effect Size From Meta-Analyses

Achievement measure	Number of effect size estimates	Mean effect size	*SD*
Elementary school	32	0.23	0.21
Lower elementary (Grades 1–3)	19	0.23	0.18
Upper elementary (Grades 4–6)	20	0.22	0.18
Middle school	27	0.27	0.24
High school	28	0.24	0.15

Note. Each effect size estimate contributing to these statistics is itself a mean effect size averaged over the studies included in the respective meta-analyses. Weighted means and standard deviations are shown, weighted by the number of studies on which each effect size estimate is based.
Sources. Compiled by the authors from 76 existing research reports and publications (Bloom et al., 2007b; Lipsey et al., 2007).

different effect size estimates even when the interventions and samples are similar.

- The usefulness of these empirical benchmarks depends on the degree to which they are drawn from high-quality studies that provide valid estimates of intervention effects and the degree to which they summarize effect sizes from similar types of interventions, populations, and outcome measures.

SUMMARY: USE EMPIRICAL BENCHMARKS TO INTERPRET EFFECT SIZES *IN CONTEXT*

Tests of the statistical significance of intervention effects follow a formal process that is well documented and widely accepted. However, the process of interpreting program impacts in terms of their policy relevance or substantive significance does not benefit from such theory or norms. If there is any norm, it is to refer to Cohen's (1988) rules of thumb for small, medium, and large effect sizes.

We argue that any such rules of thumb ignore the context that produces a particular estimate of program impact and that better guidance for interpreting impact estimates can be obtained from empirical benchmarks. We illustrate this point with three types of benchmarks: those based on normative change, those based on policy-relevant gaps, and those based on impact findings from previous research. Although each source provides a different lens for viewing a particular effect size from a particular study, all point to the importance of interpreting the magnitude of an intervention effect *in context*: of the intervention being studied, of the outcomes being measured, and of the samples or subgroups being examined. Indeed, it is often useful to use multiple benchmarks when assessing the observed impacts of an intervention. When it comes to such findings, we thus conclude that one effect size rule of thumb does not and cannot fit all.

REFERENCES

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2006a, June). *Effect sizes in education research: What they are, what they mean, and why they're important*. Presented at the Institute of Education Sciences 2006, Research Conference, Washington, DC.

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2006b, April). *Interpreting effect size findings in education research*. Presented at Department of Educational Psychology, University of Wisconsin–Madison.

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2007a, January). *Effect sizes in education research: What they are, what they mean, and why they're important*. Presented to Abt Associates Inc., Bethesda, MD.

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2007b, March). *Using empirical benchmarks for interpreting effect sizes*. Presented at the Interagency Roundtable Meeting on "The Application of Effect Sizes in Research on Children and Families: Understanding Impacts on Academic, Emotional, Behavioral, and Economic Outcomes," Washington, DC.

Bloom, H., Hill, C., Black, A., & Lipsey, M. (in press). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*.

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

Dynarski, M., Gleason, P., Rangarajan, A., & Wood, R. (1998). *Impacts of dropout prevention programs final report: A research report from the School Dropout Demonstration Assistance Program Evaluation*. Washington, DC: Mathematica Policy Research.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490–499.

Kane, T. J. (2004, January 16). *The impact of after-school programs: Interpreting the results of four recent evaluations*. William T. Grant Foundation Working Paper. Retrieved July 16, 2008, from http://www.wtgrantfoundation.org/usr_doc/After-school_paper.pdf

Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reforms? *Teachers College Record*, *110*, 1613–1640.

Lipsey, M., Bloom, H., Hill, C., & Black, A. (2007, February 6). *How big is big enough? Achievement effect sizes in education*. Presented at University of Pennsylvania Graduate School of Education.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, *26*(1), 33–51.

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

Snipes, J. C., Holton, G. I., & Doolittle, F. (2006). *Charting a path to graduation: The effect of Project GRAD on elementary school student outcomes in four urban school districts*. New York: MDRC.