MDRC Working Papers on Research Methodology

Empirical Benchmarks for Interpreting Effect Sizes in Research

Carolyn J. Hill
Howard S. Bloom
Alison Rebeck Black
Mark W. Lipsey

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

July 2007

Acknowledgments

This working paper is part of a series of publications by MDRC on alternative methods of evaluating the implementation and impacts of social and educational programs and policies. The paper will be published as an article in the forthcoming *Child Development Perspectives* journal by the Society for Research in Child Development. Many thanks are due Larry Hedges for his helpful input.

This working paper was supported by the Institute of Education Sciences in the U.S. Department of Education; the Judith Gueron Fund for Methodological Innovation in Social Policy Research at MDRC, which was created through gifts from the Annie E. Casey, Rockefeller, Jerry Lee, Spencer, William T. Grant, and Grable Foundations; and the William T. Grant Foundation.

The findings and conclusions in this paper do not necessarily represent the official positions or policies of the funders.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: Alcoa Foundation, The Ambrose Monell Foundation, Bristol-Myers Squibb Foundation, and The Starr Foundation. In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, The Sandler Family Supporting Foundation, and The Stupski Family Fund, as well as other individual contributors.

For information about MDRC and copies of our publications, see our Web site: www.mdrc.org..

Abstract

In this article we argue that 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. We apply this approach to the assessment of effect size measures for educational interventions designed to improve student academic achievement. Three types of empirical benchmarks are presented: (1) normative expectations for growth over time in student achievement; (2) policy-relevant gaps in student achievement, by demographic group or school performance; and (3) effect size results from past research for similar interventions and target populations. Our analysis draws from a larger ongoing research project that is examining the calculation, interpretation, and uses of effect sizes measures in education research. The more general message — that effect sizes should be interpreted using relevant empirical benchmarks — is applicable to any policy or program area, however.

Contents

List of Tables

Introduction

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., Magnusen et al., 2007), whether curricular interventions increase reading or mathematics achievement (e.g., Snipes et al., 2006), or whether after-school programs reduce dropout from high school (e.g., Dynarski et al., 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 policymakers 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.[1] Typically, these effect size magnitudes have been interpreted based on rules of thumb suggested by Jacob Cohen (1988), whereby an effect size of about 0.20 is considered "small"; about 0.50 is considered "medium"; and about 0.80 is considered "large." The Cohen guidelines are only broad generalizations, however, covering many types of interventions, target populations, and outcome measures. Nevertheless, it has been standard practice for researchers and policymakers to interpret effect size estimates using these guidelines.

In this article, 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 illustrate this point with three types of benchmarks: (1) normative expectations for change, (2) policy-relevant performance gaps, and (3) 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 sizes measures in education research.[2] 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.

[1] Unlike tests of statistical significance, which are influenced by sample size as well as effect magnitude, standardized effect size measures are independent of sample size.

[2] We explore two additional empirical benchmarks in our larger study: (1) an intervention's effect on rates of attainment of a performance criterion and (2) its effect or benefits relative to its costs.

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 with 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 twelfth grade for the national norming samples of seven major standardized tests in reading and six tests in math.[3] We used each test's technical manuals to obtain its mean scale score and standard deviation, by grade. (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 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.[4] 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 school, plus any learning loss experienced during the summer.[5]

The resulting growth trajectories for reading and math are shown in the "Mean" columns of Table 1. The margin of error (for a 95 percent confidence interval) 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 standard deviation. Because the margin of error for this estimate is 0.10, the lower bound of its 95 percent confidence interval 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 for grades 1 to 2 is 0.97 standard deviation; for grades 5 to 6 it is 0.32 standard deviation; and for grades 11 to 12 it is only 0.06 standard deviation. While 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. The same

[3] California Achievement Test - 5th edition (CAT5); Stanford Achievement Test Series - 9th edition (SAT9); TerraNova-Comprehensive Test of Basic Skills (CTBS); Gates-MacGinitie; Metropolitan Achievement Test (MAT8); TerraNova-California Achievement Tests (CAT); and Stanford Achievement Test Series: 10th Edition (SAT10). The Gates-MacGinitie does not include a mathematics component.

[4] The weighting formula was drawn from Hedges (1982).

[5] 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 the same as those calculated from cross-sectional data (grade-to-grade differences for a given year), except for the transition from ninth grade to tenth grade, when large numbers of students drop out of school. (Results are available from the authors on request.)

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	*(+/- 0.22)*
Grade 1 - 2	0.97	*(+/- 0.10)*	1.03	*(+/- 0.11)*
Grade 2 - 3	0.60	*(+/- 0.10)*	0.89	*(+/- 0.12)*
Grade 3 - 4	0.36	*(+/- 0.12)*	0.52	*(+/- 0.11)*
Grade 4 - 5	0.40	*(+/- 0.06)*	0.56	*(+/- 0.08)*
Grade 5 - 6	0.32	*(+/- 0.11)*	0.41	*(+/- 0.06)*
Grade 6 - 7	0.23	*(+/- 0.11)*	0.30	*(+/- 0.05)*
Grade 7 - 8	0.26	*(+/- 0.03)*	0.32	*(+/- 0.03)*
Grade 8 - 9	0.24	*(+/- 0.10)*	0.22	*(+/- 0.08)*
Grade 9 - 10	0.19	*(+/- 0.08)*	0.25	*(+/- 0.05)*
Grade 10 - 11	0.19	*(+/- 0.17)*	0.14	*(+/- 0.12)*
Grade 11 - 12	0.06	*(+/- 0.11)*	0.01	*(+/- 0.11)*

SOURCES: Annual gain for reading is calculated from seven nationally normed tests: CAT5, SAT9, TerraNova-CTBS, 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

pattern of findings was observed for tests of social studies and science. (Results are available from the authors on request.)

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 for upper grades. Second, the sample composition for upper grades is changing across grades due to 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 sub-groups of students.

Nevertheless, because the preceding patterns 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 — e.g., an effect size of 0.10 standard deviation — would constitute a relatively smaller substantive change for students in early grades than for students in later grades.[6] Thus, it is important to interpret a study's effect size estimate in the context of natural growth for its target population.

- 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 (2005) 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, to eliminate the achievement gaps between Black and White, rich and poor, and males and females . . . it is natural to evaluate reform effects by comparing them to the size of the gaps they are intended to ameliorate" (Konstantopoulos and Hedges, 2005, p. 4). 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). Gaps in reading and math scores are presented by race/ethnicity, income (free/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 standard deviation lower than white fourth-graders on the reading assessment, and they scored 0.99 standard deviation lower on the math assessment. A gap between blacks and whites is observed at each of the three grade levels, though is somewhat smaller in twelfth 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 to score higher in math. These gender gaps are typically much smaller than corresponding race/ethnicity or income gaps.

[6] This point does *not* imply that it is necessarily *easier* to produce a given effect size change — e.g., of 0.10 — for early grades than for later grades.

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.

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 (for example, 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 — e.g., 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 for others (e.g., that for males and females).[7] Thus, it is important to interpret a study's effect size estimate in the context of the subgroups of interest.

- 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 impor-

[7] This point does *not* imply that it is necessarily *easier* to produce a given effect size change — e.g., of 0.10 — to close the gaps for some groups than for others.

tant to interpret a study's effect size estimate in relation to the policy-relevant gap for a particular outcome measure and target population.

- Additionally, performance gaps can provide a relevant benchmark for effect sizes from interventions even if they are not explicitly intended to reduce a particular performance gap.

In addition to gaps based on student characteristics, performance gaps among schools may also be relevant for policy. As Konstantopoulos and Hedges 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" (2005, p. 4).

Table 3 illustrates these kinds of gaps. However, instead of comparing low-performing schools with high-performing schools, the table 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, the adjusted mean performance over a two- or three-year period is calculated.[8] The distribution of average school performance for each grade in each district is then generated. 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).[9]

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 standard deviation. 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.

[8] Means are regression-adjusted for test scores in prior grade and students' demographic characteristics (race/ethnicity, gender, age, and free-lunch status). Performance is measured using nationally normed standardized tests: for District I, scale scores from the 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.

[9] 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.

Table 3

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

	District Findings			
	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

SOURCES: ITBS for District I, SAT9 for District II, MAT for District III, and SAT8 for District IV. See description in text for further details on the sample and calculations.

NOTE: "NA" indicates that a value could not be computed due to missing test score data. Means are regression-adjusted for test scores in prior grade and students' demographic characteristics.

Findings in Table 3 illustrate that:

- A particular effect size (e.g., 0.10 standard deviation) 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 (e.g., 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 with those from previous studies for similar grade levels, interventions, and outcomes? This approach uses results from research synthesis, or meta-analysis. We illustrate it with two such analyses.

The first analysis summarizes estimates of achievement effect sizes from random assignment studies of educational interventions. These results are thus based on the most rigorous impact design available.[10] 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.[11] Because most studies report multiple effect size estimates (e.g., for multiple outcomes or grades), a total of 468 effect sizes 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 standard deviation; the corresponding mean effect size for middle schools is 0.51 standard deviation, and that for high schools is 0.27 standard deviation. 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, when interpreting these results, it should be noted that the

[10] See Shadish, Cook, and Campbell (2002) for a discussion of random assignment experiments.

[11] The research synthesis did not include random assignment studies of special education students or of students with clinical problems; nor did 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 percent.

Table 4

Summary of Effect Sizes from Randomized Studies

Achievement Measure	Number of Effect Size Estimates	Mean Effect Size	Standard Deviation
Elementary schools	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 schools	36	0.51	0.49
High schools	43	0.27	0.33

SOURCES: Compiled by the authors from 61 existing research reports and publications (reporting on 95 independent subject samples).

NOTE: Unweighted means are shown across all effect sizes and samples in each category.

interventions and target populations being studied might also differ across the three categories of outcome measures used.

Our second example of an empirical benchmark from a research synthesis is a “meta-analysis of meta-analyses.” These findings summarize the results of 76 meta-analyses of past studies of educational interventions in kindergarten through twelfth grade that reported mean achievement effect sizes for experimental and quasi-experimental studies and that provided some breakdown by grade range (elementary, middle, high school).[12]

Descriptive statistics from this meta-analysis of meta-analyses are reported in Table 5. Averaged over the many different interventions, studies, and achievement outcomes encompassed in these meta-analyses, the mean effect sizes are in the 0.20 to 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.

[12] A total of 192 meta-analyses of educational interventions were located. 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, a weighted mean was computed (weighting by the number of studies included in each meta-analysis).

Table 5

Distributions of Mean Effect Sizes from Meta-Analyses

Achievement Measure	Number of Effect Size Estimates	Mean Effect Size	Standard Deviation
Elementary school	32	0.23	0.21
Lower elementary (1-3)	19	0.23	0.18
Upper elementary (4-6)	20	0.22	0.18
Middle school	27	0.27	0.24
High school	28	0.24	0.15

SOURCES: Compiled by the authors from 76 existing research reports and publications.

NOTES: 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.

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, achievement — may result in 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 and the degree to which they summarize effect sizes with regard to similar types of interventions, target 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.

In this article, 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. While 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

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Dynarski, Mark, Philip Gleason, Anu Rangarajan, and Robert Wood. 1998. *Impacts of Dropout Prevention Programs Final Report: A Research Report from the School Dropout Demonstration Assistance Program Evaluation.* Washington, DC: Mathematica Policy Research, Inc.

Hedges, Larry V. 1982. "Estimation of Effect Size from a Series of Independent Experiments." *Psychological Bulletin* 92 (2): 490-499.

Kane, Thomas J. 2004. "The Impact of After-School Programs: Interpreting the Results of Four Recent Evaluations." William T. Grant Foundation Working Paper (January 16). Web site: http://www.wtgrantfoundation.org/usr_doc/After-school_paper.pdf.

Konstantopoulos, Spyros, and Larry V. Hedges. 2005. "How Large an Effect Can We Expect from School Reforms?" Working Paper 05-04. Evanston, IL: Northwestern University, Institute for Policy Research.
Web site: http://www.northwestern.edu/ipr/publications/papers/2005/WP-05-04.pdf.

Magnuson, Katherine A., Christopher Ruhm, and Jane Waldfogel. 2007. "Does Prekindergarten Improve School Preparation and Performance?" *Economics of Education Review* 26 (1): 33-51.

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

Snipes, Jason C., Glee Ivory Holton, and Fred Doolittle. 2006. *Charting a Path to Graduation: The Effect of Project GRAD on Elementary School Student Outcomes in Four Urban School Districts.* New York: MDRC.