Some Food for Thought
About Effect Size

Howard S. Bloom
MDRC
Mark W. Lipsey
Vanderbilt University
March 11, 2004

The following examples highlight some important facts of life about the metric of effect size. The definition of effect size that is used equals the difference in mean outcomes for two groups (usually a treatment group and a control group) divided by the standard deviation of the outcome measure for a relevant group (usually the control group). We will use these examples in our discussion to illustrate how one should think about the precision that is needed for a study of program impacts and thus, the sample size and allocation that is required to produce this precision. The examples are presented in no particular order and do not lead to a specific conclusion that applies to all situations. Rather they are intended to illustrate the range of situations that might exist and the corresponding range of considerations that must be taken into account when designing an experimental study of the effects of an intervention.

1. The Effect of Aspirin on Heart Attacks[1]

In 1987 a randomized clinical test of aspirin's ability to prevent heart attacks was ended prematurely because the observed reduction in heart attacks produced by aspirin was too large to justify continuing to give a placebo to control group members. This test was based on a sample of 22,071 physicians, half of whom were randomized to an aspirin regimen and half of whom were randomized to a placebo. 1.71 percent of the physicians who received the placebo experienced a heart attack, whereas only 0.94 percent of the physicians who received the aspirin regimen did so. Hence, aspirin reduced the rate of heart attacks by 0.77 percentage points. This represents an effect size of 0.06 standard deviations.

Question: *Why was this effect size large enough to stop the test and prompt a declaration of victory for aspirin? What other way/s might this impact be presented?*

[1] Rosenthal, Robert (1994) "Parametric Measures of Effect Size" in Harris Cooper and Larry V. Hedges, ***The Handbook of Research Synthesis*** (New York: Russell Sage Foundation).

2. Cohen's Rules of Thumb[2]

Jacob Cohen is renowned for promulgating the characterization of effect sizes of roughly 0.2 standard deviations as "small", 0.5 standard deviations as "medium" and 0.8 standard deviations as "large". Cohen (pp. 532) states that: "these proposed conventions were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible. The values chosen had no more reliable a basis than my own intuition. They were offered as conventions because they were needed in a research climate characterized by a neglect of attention to issues of magnitude. The ES measures and conventions have been successful, widely adopted not only for power analysis, but more widely, for example, in ES surveys and in meta-analyses. But there are difficulties and much room for misunderstanding."

Question: *Why has Cohen's characterization been used so widely by social scientists for the past several decades and what does this mean for future research on high school programs?*

3. Lipsey's Empirical Benchmarks[3]

Mark Lipsey has provided an empirical basis for judging the magnitudes of effect sizes from studies of psychological, educational and behavioral treatments. His findings represent the distribution of mean effect size estimates from 102 meta-analyses, which summarize the results of 6700 individual studies (most of which were nonexperimental) involving almost 800,000 subjects. The bottom third of this distribution ranges from 0.00 to 0.32 standard deviations, the middle third ranges from 0.33 to 0.55 standard deviations, and the top third ranges from 0.56 to 1.20 standard deviations. The midpoints of these three categories are 0.15, 0.45 and 0.90 standard deviations, respectively. This is astonishingly consistent with Cohen's rules of thumb.

Question: *Does this mean that Cohen's rules of thumb are a good basis for judging the magnitudes of effects for all programs? For high school programs?*

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

[3] Lipsey, Mark W. (1990) ***Design Sensitivity: Statistical Power for Experimental Research*** (Newbury Park, CA: Sage Publications).

4. The Effect of Career Academies on Future Earnings[4]

Career Academies, which were established more than 30 years ago as an alternative to large comprehensive high schools, now exist in more than 2,500 locations across the U.S. Some of the key distinguishing features of Career Academies are that they: combine academic and technical instruction that is focused on career themes, build partnerships with employers that provide linkages for students, and offer many types of work-based learning opportunities. Since 1993 MDRC has been conducting a randomized experimental study of Career Academies in nine high schools from around the Country. The most recent results from this study examine labor market impacts during the first four years after high school.[5] Although little or no impacts were observed for young women in the study sample, the impacts observed for young men were pronounced. Career Academies increased their average earnings during the follow-up period by $212 per month. This is 18 percent higher than their earnings would have been without the program. And it is more than twice the roughly $100 monthly earnings difference that exists between young workers with one or two years of post-secondary education and those with only a high school diploma or a GED.[6] When expressed as an effect size, the observed impact equals 0.30 standard deviations.

Question: *What utility, if any, does an effect size metric have for this outcome measure and why? What should one conclude from assessing the observed impact in the several ways presented?*

[4] Based on computations for Kemple, James (forthcoming) “Career Academies: Impacts on Labor Market Outcomes and Educational Attainment (New York: MDRC).

[5] More precisely, they focus on labor market outcomes during the first four years after what would be the date of on-time graduation for sample members.

[6] Pond, Nathan, Andrew Sum, Turub’sky, Mykhaylo and Frank Meredith (2002) “Trends in the Level and Distribution of the Weekly and Annual Earnings of Young Adult Men and Women in the U.S., 1973-2001” (Washington, DC: National League of Cities Institute on Youth, Education and Young Families).

5. The Longer-term Effects of Small Classes[7]

Project STAR (Student/Teacher Achievement Ratio), the Tennessee Class Size Experiment, is widely hailed as providing the most compelling evidence that exists on the effect of class size on student achievement. This landmark study randomized teachers and students in kindergarten through third grade from 79 schools located in 42 Tennessee school districts to either a smaller class, with 13 to 17 students, or a larger class, with 22 to 26 students. Depending on when students entered the schools in the study, they spent between one and four years in either a smaller class or a larger class (with some students shifting between these alternatives). Follow-up data for these students made it possible to measure the impacts of class size in grades K – 3 on their performance in reading, mathematics and science tests in grades 4, 6 and 8. These follow-up data indicate that: "the average effect of small classes was statistically significant and positive for both mathematics and reading achievement at every grade level, ranging from 0.11 to 0.20 standard deviation units. The small class effect was positive for science achievement at all grades (ranging from 0.10 to 0.17 standard deviation units) and was statistically significant for both Grades 6 and 8. There was little evidence of interaction between gender and class size ……….. there was no evidence that small class effects varied across schools" (p. 132).

Question: *These effects are statistically significant, but are they of practical significance? What else would you want to know before deciding whether this is an educationally meaningful effect?*

[7] Nye, Barbara, Larry V. Hedges and Spyros Konstantopoulos (1999) "The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment," ***Educational Evaluation and Policy Analysis,*** 21:2, 127-142.

6. The Black-White Test Score Difference[8]

Controversy has raged for decades over the magnitude, the causes and the consequences of the difference between standardized test scores for Black students and white students. The table below reports this difference for national samples of seventeen-year-old students in selected years between 1971 and 1994. This information, which was obtained from the National Assessment of Educational Progress (NAEP), is reported as an effect size. Effect size was defined as the mean score for Blacks minus the mean score for whites divided by the standard deviation of individual scores for students of all races, adjusted for measurement error.

Black-White Differences in Mean NAEP Scores
For Seventeen-Year-Old Students

Year	**Mean Score Difference in Standard Deviations**	
	Reading	**Mathematics**
1971	- 1.15	--
1975	- 1.19	--
1978	--	- 1.07
1980	-1.19	--
1982	--	- 0.98
1984	-0.79	--
1986	--	- 0.93
1988	-0.55	--
1990	-0.72	- 0.68
1992	-0.86	- 0.87
1994	-0.66	- 0.89

SOURCE: Table 5-2 of Hedges and Nowell (1998)
NOTE: Standard deviations were computed for students of all races and adjusted for measurement error.

***Question:** What are at least two ways that this information could be used to assess impact estimates for high school reforms? What is implied by the fact that effect size is adjusted for measurement error?*

[8] Hedges, Larry V. and Amy Nowell (1998) “Black-White Test Score Convergence Since 1965,” in Christopher Jencks and Meredith Philips, editors, ***The Black-White Test Score Gap*** (Washington, DC: Brooking Institution Press).

7. The Impact of Welfare-to-Work Programs on Earnings[9]

MDRC recently conducted a quantitative research synthesis of three of its largest multi-site welfare-to-work experiments: California's Greater Avenues for Independence (GAIN) program, Florida's Project Independence (PI), and the National Evaluation of Welfare-to-Work Strategies (NEWWS). Data for this analysis represent random assignment experiments in 59 local welfare offices from seven states. These data reflect the experiences of 69,399 female single parents who had applied for, or were receiving welfare when they were randomized to a program group, which was offered a wide range of special employment and training services, or a control group, which only had access to standard services provided by existing welfare systems. A separate program impact was estimated for each of the 59 local program offices. The table below presents a summary of the findings obtained for impacts on average total earnings during the first two years after random assignment. This summary reports the 25th, 50th and 75th percentile values for the 59 impacts in three different metrics: (1) total constant dollars, (2) a percentage of what earnings would have been without the program (its counterfactual), and (3) an effect size in units of the standard deviation of the outcome measure for individual control group members.

Impacts of Welfare-to-Work Programs
On Average Two-Year Follow-up Earnings

Percentile	Impact in Dollars	Impact in Percent	Impact in Standard Deviations
25th	42	< 1	0.00
50th	738	17	0.09
75th	1,615	36	0.17

SOURCE: Computations by Bloom, Hill and Riccio (2003)

***Question:** What do these findings suggest about the use of effect size for reporting the impacts of welfare-to-work programs? What do the findings suggest about interpreting effect sizes from research on high schools?*

[9] Bloom, Howard S., Carolyn J. Hill and James A. Riccio (2003) "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments", ***Journal of Policy Analysis and Management***, 22:4, 551-575.