Paper 2: How should we calculate effect sizes? What are common mistakes in the calculation of effect sizes? How does research design influence the calculation of effect sizes?
Dr. Howard Bloom, Senior methodologist at MDRC

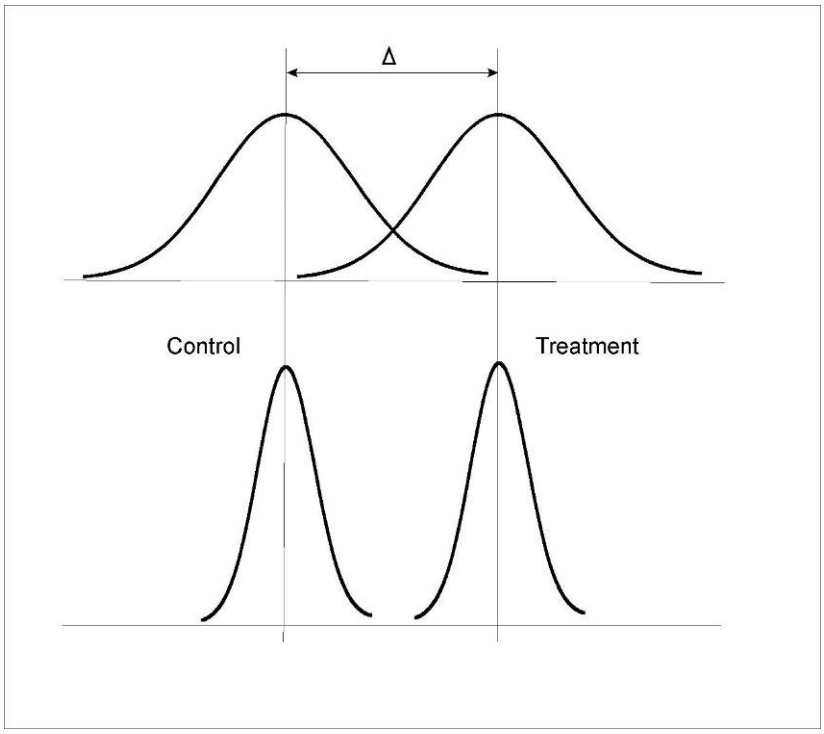
There are two considerations when computing or interpreting Effect Sizes: (a) different definitions of effect size convey different impressions and (b) interpretations of the magnitudes of effect sizes depend on their context.

The Standardized Mean Difference Effect Size (ES)

$$ES = \frac{\bar{Y}_E - \bar{Y}_C}{\sigma}$$

$$ES = \frac{410 - 400}{50} = 0.20\sigma$$

The standardized mean difference effect size is a relative concept that depends upon the standard deviation used. Dividing the difference by a large standard deviation, results in a relatively small standardized mean difference (top curve) whereas dividing the difference by a relatively small standard deviation, results in a larger effect size (bottom curve).



Variance Components Framework

A central tenet of effect sizes is understanding that the choice of standard deviation used to calculate an effect size influences its magnitude and thus its meaning. This point can be illustrated using a variance components framework in terms of test scores. By definition, there is some overall standard national variance in individual test scores. The total national variance can be decomposed into the variations across state, and within state across various districts, within

districts across schools, subgroups of students, students within groups and then each one of these test scores is a measure of some latent achievement variable. With all of these variances, different researchers will choose one or a combination of variances for their study and provide justification for their choice. Then, perhaps another researcher will perform a meta-analysis, but the effect sizes are not comparable unless the same metric is applied.

$$\sigma^2_{U.S.} = \sigma^2_{state} + \sigma^2_{district} + \sigma^2_{school} + \sigma^2_{subgroup} + \sigma^2_{student} + \sigma^2_{error}$$

$$\sigma_{U.S.} = \sqrt{\sigma^2_{state} + \sigma^2_{district} + \sigma^2_{school} + \sigma^2_{subgroup} + \sigma^2_{student} + \sigma^2_{error}}$$

Here is an example of student-level versus school-level standard deviations:

$$\sigma_{student} = \sqrt{\tau^2 + \sigma^2} \qquad \tau^2 = between - school - \text{var}\,iance$$

$$\sigma_{school} = \sqrt{\tau^2_2 + \frac{\sigma^2}{n}} \qquad \sigma^2 = within - school - \text{var}\,iance$$

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} = ICC$$

The table below depicts the ratio of student-level to school-level standard deviations. As intra-class correlations increase, schools are more different from each another on average. For example, if you have 50 third graders per school and the intra-class correlation is .20, the ratio of the student-level standard deviation to the school-level standard deviation is 2.15. In other words, the student-level standard deviation is a little over twice the size of the school-level standard deviation. Depending on the standard deviation used, the magnitude of the effect size varies greatly.

Students in a grade per school (n)	**Intra-class correlation (ρ)**		
	.05	.10	.20
50	3.81	2.91	2.15
100	4.10	3.03	2.19
200	4.27	3.09	2.21
400	4.37	3.13	2.22

Adjusted versus Unadjusted Standard Deviation

Another factor to consider is whether to use an unadjusted standard deviation or a *regression*-adjusted standard deviation. Once again, the choice in standard deviation can result in effect sizes that look very different.

$$\sigma^2_{unexplained} = (1 - R^2)\sigma^2_{total}$$

R^2	Ratio of unadjusted to adjusted standard deviations
0.1	1.05
0.3	1.20
0.5	1.41
0.7	1.83
0.9	3.16

The following table provides a sense of the calculations and the implications of using an unadjusted or a *reliability*-adjusted standard deviation.

$$\sigma^2_{true} = \lambda \sigma^2_{observed}$$

Reliability	Ratio of unadjusted to adjusted standard deviations
0.9	**1.05**
0.7	**1.20**
0.5	**1.41**
0.3	**1.83**
0.1	**3.16**

Assessing and Interpreting an Effect Size

The research community should develop better conventions and contingencies for understanding effect sizes. This effort will assist the field in proceeding in a more orderly manner, require researchers to report their methodology and justification, and give the reader the ability to review the same findings.

The context of "when" and the term "how" can assist in the calculation and discussion of effect sizes. For example, effect size calculations should be considered when designing an intervention study and determining the level of precision needed; when interpreting the results of an intervention study; when assigning adjectives in the final report; and when synthesizing intervention studies. Another way to discuss effect sizes can be discussed in terms of “how.” For example, comparing the external criterion or standard to either a related outcome construct or a related context.

Cohen and Lipsey offer the prevailing guidelines for interpreting an effect size

Cohen
Small = 0.20 s
Medium = 0.50 s
Large = 0.80 s

Lipsey
Small = 0.15 s
Medium = 0.45 s
Large = 0.90 s

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

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

A common guideline for gauging achievement effects in education is an effect size equal to, or greater than, .25. This level has been defined as "educationally significant." In Tallmadge's *The Joint Dissemination Review Panel IDEABOOK (1977)* he stated, "One widely applied rule is that the effect must equal or exceed some proportion of a standard deviation– usually one-third, but at times as small as one-fourth– to be considered educationally significant" (p. 34).

Conclusion

When interpreting the magnitudes of effect sizes, "one size" does not fit all. Instead, researchers should interpret magnitudes of effects *in context* of the interventions being studied, of the outcomes being measured, and of the samples/subsamples being examined. Consider different frames of reference in context, instead of a universal standard such as external criteria, normative change, policy-relevant gaps, observed effect size distributions, or intervention costs. As a parting word of caution, standardized effect size measures should only be used when necessary. For example, when the original outcome measure does not have a meaningful metric, or when comparing intervention effects that are measured in different metrics.