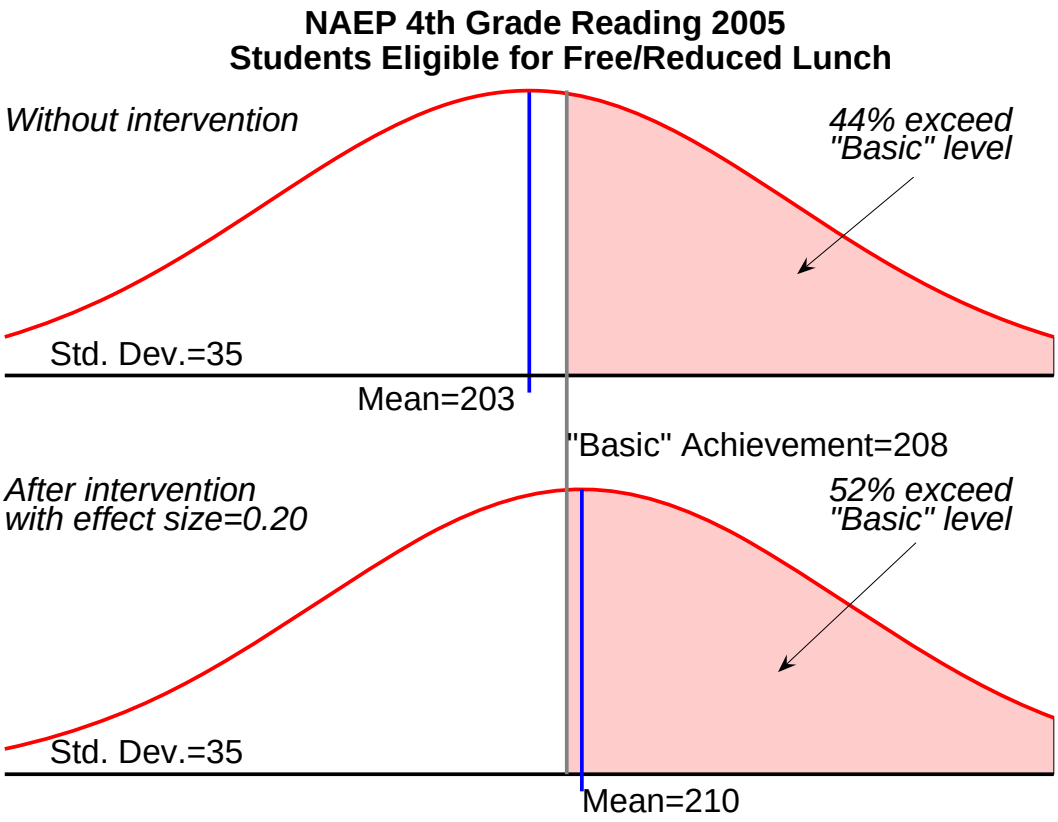
Panel 2: Interpreting and Contextualizing Effect Sizes

Belinda Sims, Health Scientist administrator at the National Institute of Drug Abuse, served as the moderator for Panel 2. She introduced the three speakers: Carolyn Hill, Margaret Burchinal, and Hendricks Brown.

Paper 1: Using Empirical Benchmarks for Interpreting Effect Sizes.
Carolyn Hill, Georgetown University (joint work with Howard Bloom (MDRC), Alison Black (MDRC), and Mark Lipsey (Vanderbilt)

This paper presents some preferred approaches for assessing effect sizes in context. For example, effect sizes for K-12 education could be interpreted by comparing the effect size from the study with: (a) attainment of a performance criterion; (b) normative expectations for change; (c) policy-relevant performance gaps; and (d) effect size distributions from similar studies. Intervention costs are another consideration when interpreting effect sizes, but not included in this presentation.

<u>Attainment of a Performance Criterion</u>
The following chart utilizes information from the National Assessment of Educational Progress (NAEP) to illustrate how an effect size might be interpreted in contexts where external performance criteria are relevant. NAEP has three externally defined achievement levels of Basic, Proficient, and Advanced. The normal curve at the top shows the distribution of the outcomes without intervention. The mean scale score for 4^{th} graders who were eligible for a free/reduced-priced lunch in 2005 was 203, with a standard deviation of .35. This curve also marks the "Basic" level of achievement (scale score = 208, shown by the gray line). Approximately 44% of 4th graders exceeded the "Basic" level.



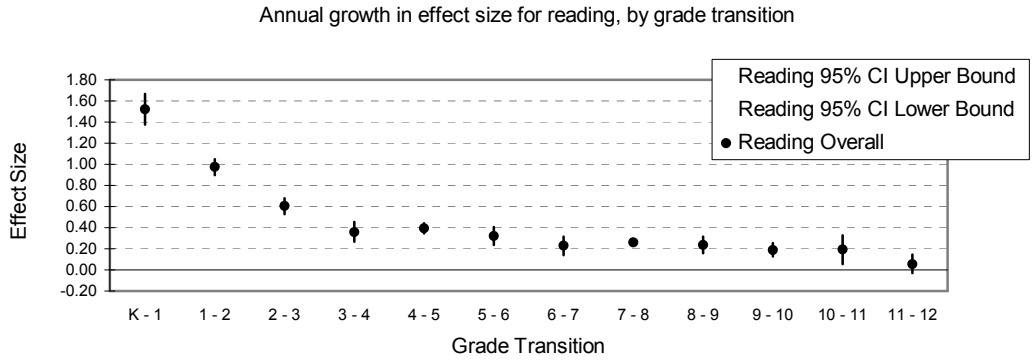
Next, consider an invention with an effect size of 0.20, which would be considered “small” by the Cohen’s guidelines. This would shift the curve to the right and raise the mean level of achievement to 210. In this instance, about 52% of these students would exceed the “Basic” level of achievement by an increase of 8 percentage points from the non-intervention state. In the context of a particular study that targets low-income children, but uses a different outcome measure, the magnitude of the effect might be interpreted in relation to this increase in “Basic” achievement from the NAEP.

Normative Expectations for Change

Another empirical benchmark might be normative expectations for change or some measure of natural growth. This can be illustrated by using estimated annual gains in effect size from national norming samples for standardized tests. Up to seven tests were used for reading, math, science, and social science. The mean and standard deviation of scale scores for each grade were obtained from test manuals and the standardized mean difference across succeeding grades was computed. These results were averaged across tests and weighted according to Hedges (1982).

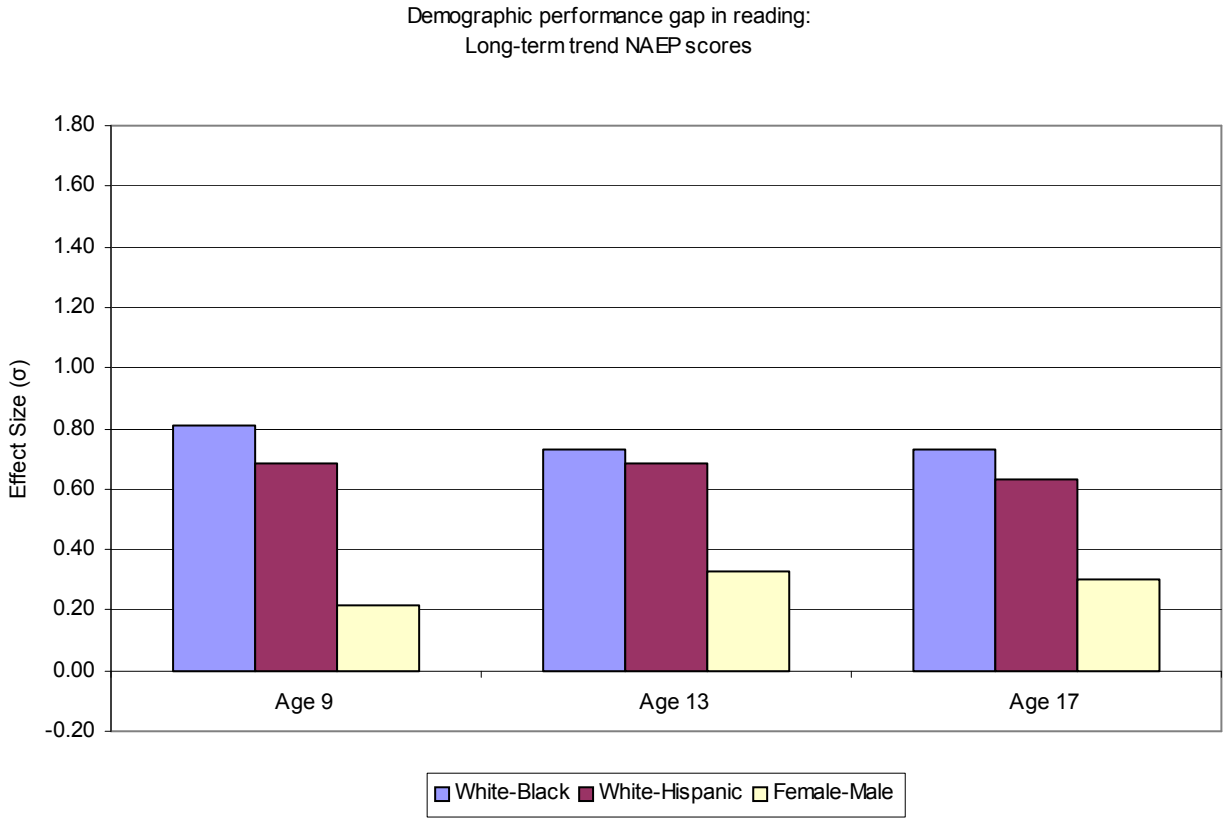
The following chart shows the natural growth over a year in effect size for reading. This distribution is striking in that the effect size declines across the years. An effect size of 0.20 in the lower elementary range constitute a relatively small change compared to the natural growth over that period. An effect size of 0.20 in the other grades is relatively large in contrast. So in thinking about effect sizes in context, this example shows that the grade may provide an important context. Even across subject matters, different magnitudes in effect are apparent. This

emphasizes the importance of looking at different outcomes measures within a grade, in addition to looking at growth across different grades.



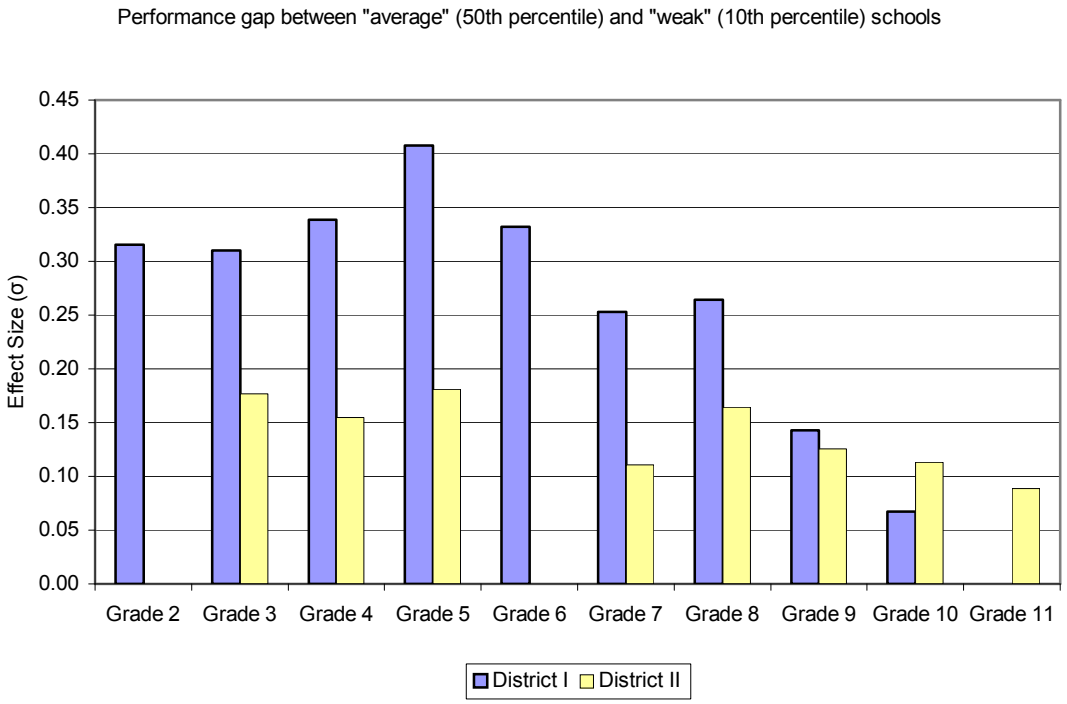
Policy-Relevant Performance Gaps

Another type of benchmark would be policy-relevant performance gaps. One type is demographic performance gaps from selected tests. Effectiveness of interventions can be judged relative to the sizes of existing gaps across demographic groups and effect size gaps may vary across grades, years, tests, and districts. The following chart illustrates an example of three different types of demographic performance gaps in reading from NAEP.



It is important to note that an effect size of 0.10 or 0.20 may be interpreted differently depending on which performance gaps it is being interpreted against. Another note is that race/ethnicity and gender gaps may look different at another grade level or for different types of tests. Even across grades, the magnitudes of the difference between race/ethnicity gaps to the gender gaps are quite different. Recognizing and understanding these factors can help explain why different types of patterns can be expected.

Another performance gap of interest might be for the same type of students in different schools. Using this approach, the researcher should estimate a regression model that controls for student characteristics: race/ethnicity, prior achievement, gender, overage for grade, and free lunch status. Then, infer performance gap (in effect size) between schools at different percentiles of the performance distribution. The following chart is an example of difference between the "average" schools and "weak" schools.



Effect Size Distributions from Similar Studies

Another type of empirical benchmark is effect size distributions from similar studies. An example is the distribution of achievement effect sizes from 421 random assignment studies of education interventions (M = 0.41, SD = 0.47). The effect sizes are associated with various percentiles. The median effect size for the total sample of studies was 0.34. When establishing a benchmark from similar studies, it is not clear that this overall distribution (which includes different types and levels of intervention, etc.) would be the preferred type to refer to. Instead, one might be interested in breakdowns by type of achievement measures and grade level:

Achievement Measure	n	Mean	SD
Standardized Test (Broad)			
Elementary	25	0.10	0.30
Middle	3	0.06	0.36
Standardized Test (Narrow)			
Elementary	115	0.31	0.42
Middle	12	0.41	0.33
Specialized Topic/Test			
Elementary	204	0.53	0.51
Middle	19	0.48	0.44
High	40	0.33	0.35

Conclusion

When interpreting the magnitudes of effect sizes, “one size” does not fit all. Instead, interpret magnitudes of effects in the context of the interventions being studied, the outcomes being measured, and/or the sample being examined. In addition, rather than interpreting effect sizes against a universal guideline, consider performance criterion, normative change, policy-relevant gaps, observed effect size distributions, and intervention costs.