

**Panel 3: Considering effect sizes within the policy context**  
**Helping Educators, Parents and Students Find Meaning in Research Results**  
Harris Cooper, Duke University

There are difficulties in communicating results of research to audiences lacking statistical expertise. Three important issues are: (a) how do we present the strengths and limitations of research designs to audiences lacking research expertise; (b) can researchers use adjectives that are inherently qualitative, such as “significant” or “promising”, relative such as “large” or “small” to describe the size of effects; and (c) are there metrics that permit audiences lacking statistical expertise to interpret research outcomes using yardsticks that have meaning to them?

Presenting results in the context in which the research was conducted is important. When talking about research designs with audiences, it is important to make distinctions among four different kinds of research design and to use terms that most educational audiences will understand. Useful terms are “purposely manipulated research” such as experiments and quasi-experiments, “modeling research,” and “simple association.”

There are issues that arise when you use adjectives to describe the size of effects that are inherently qualitative. In social and behavioral science, when we use the term “significant effects” we mean “different from zero (with some probability)”. In common speech, significant effects means “having or conveying a meaning.” The difference between the two is that common speech refers to an effect that is of consequence while social science makes no claim about importance. In social and behavioral science when we use the term “promising” we mean the results are in a favorable direction, with qualification because there is an inability to draw casual inferences (based on less rigorous designs) and/or there is a lack of statistical power. In common speech, “promising” means “likely to turn out well.” The difference between the two is that in common speech “promising” is a prediction about the future, in social science it is an assessment of past research acknowledging both strengths and weaknesses. Few social scientists would “promise” positive future results based on “promising” past results.

In social science, when we use the term “small and large effects” researchers are often referring to Cohen (1988) small:  $d = .20$ ,  $r = .10$  whereas large:  $d = .80$ ,  $r = .50$ . In common speech, “small” means “not large, of limited size especially in comparison with others” and “large” means “ample, wide, great, of considerable or relatively great size.” The difference between the two is that while both are relative, Cohen defines small and large relative to typical effect sizes encountered throughout the social sciences, in common speech the yardstick involves the same class or kind.

There are metrics that permit audiences lacking statistical expertise to interpret research outcomes using yardsticks that have meaning to them. The standardized mean difference can be translated into class rank and grades on a curve to give meaning for general audiences. Class rank is simply a translation of the standardized mean difference. For example, homework’s effect on achievement can be described as “the average student doing homework performed better than about 73% of students doing no homework.” Grading on a curve, the student who received the middle C grade in the homework group would have moved up to a B- grade had she or he been graded in the no-homework group.

In conclusion, there is no interpretation of effect sizes without a narrative context that “translates” complex research distinctions into common language and acknowledges the strengths and limitations of the research. An appropriate system for applying interpretive labels to effect sizes likely will never be found. When effect sizes are “translated” into metrics that have meaning for general audiences, the need for labels will likely disappear.