

*Measuring Effect Sizes, The Effect of Measurement Error*

Don Boyd, PhD – SUNY Albany

Pam Grossman, PhD – Stanford University

Hamp Lankford, PhD – SUNY Albany

Susanna Loeb, PhD – Stanford University

Jim Wyckoff, PhD – University of Virginia

Wednesday, April 23

1:30-3:30pm

---

With the increasing availability of administrative databases that include student-level achievement, the use of value-added models in education research has expanded rapidly. These models allow us to explore how a wide variety of policies and measured school inputs affect the academic performance of students. From a policy perspective, a key question is whether such effects are sufficiently large to achieve various policy goals. For example, would hiring teachers having stronger academic backgrounds sufficiently increase test scores for traditionally low-performing students to offset the increased cost of doing so?

The *effect sizes* of independent variables are measured as the estimated effect of a one standard deviation change in the variable divided by the standard deviation of test scores in the relevant population of students. Effect size estimates derived from VAM models employing administrative databases are typically quite small. This is the case across a variety of different databases and applications (see Goldhaber, 2007 for a recent review of many of these studies.). As a result, it is commonly believed that observed teacher and school attributes have little effect on improving student achievement. However, these measures of effect size may underestimate the magnitude of true effects for several reasons. We explore the empirical significance of two of these in this paper.

First, we argue that the standard deviation of gain scores, not the standard deviation of scores, should be employed in calculating most effect sizes. Because education is a cumulative process, the dispersion in academic achievement at a point in time depends upon initial differences when students enter school and the pattern of home, school and other factors in succeeding years. However, we typically measure the effect of interventions lasting a single school year (e.g., a student in a particular grade having a teacher with stronger academic background). Such interventions are likely to result in students making only modest changes in their overall educational attainment. Thus, comparing the estimated effect of any short-lived intervention to the dispersion in test scores measuring knowledge and skills at a point in time misleadingly can suggest that the intervention did not meaningfully alter the trajectory of gains over the relevant period.

The second, issue concerns the need to account for test measurement error in reported effect sizes. The standard deviation of observed scores in the denominator of the effect size measure reflect such measurement error as well as the dispersion in true academic achievement of students, thus overstating variability in achievement. It is the size of an estimated effect relative to the dispersion in true achievement, or the gain in true achievement, that is of interest. Because gain scores have measurement error in pre-tests and post-tests, netting out measurement error is especially important in this context.

Adjusting estimates of effect-size to account for these considerations is straightforward if one knows the extent of test measurement error. Technical reports provided by test vendors typically only provide information regarding the measurement error associated with the construction of the test itself (e.g., a particular set of questions being selected). However, there are a number of other factors (e.g., a student having a bad day) which can result in a particular test score not accurately reflecting true academic achievement. Using test scores of students in New York City during the 1999-2007 period, we estimate the overall extent of measurement error and how test measurement error varies across students. We apply these estimates in an analysis of how various attributes of teachers affect the test-score gains of their students and find that estimated effect sizes including the two adjustments are four times larger than estimates that do not.