Estimates of External Bias in Impact Evaluations that Select Sites Purposively

Stephen H. Bell (Abt Associates), Larry L. Orr (Johns Hopkins University), Robert B. Olsen (Abt Associates), Elizabeth A. Stuart (Johns Hopkins University)

Substantial bias may result when evaluations of government programs purposively—rather than randomly—select the sites from which data will be collected and estimates derived (see Olsen, Orr, Bell, and Stuart, 2010, for a formal expression of that bias). To estimate "external validity" bias of this sort, we use data from a recent evaluation of the Reading First program covering all school districts in 15 states. From these data, we compute a benchmark estimate of Reading First's impact on student achievement in all districts in all 15 states. We then compare this benchmark to estimates of impact from the subset of districts that participated in a different education impact evaluation which selected districts purposively. The difference between the purposive sample estimate and the benchmark estimate is an estimate of the external bias that would have resulted if Reading First had been evaluated in the purposively-chosen sites alone. Measures of this bias are computed for 12 different purposive-sample evaluations taken from the education literature. The result is robust empirical evidence on the question of whether policy impact evaluations based on purposive samples have adequate external validity.