Randomized Controlled Trials (RCTs) are commonly used to estimate causal impacts of educational programs, and have been called the “gold-standard” of evaluation (Mosteller & Boruch, 2002; Rossi, Lipsey & Freeman, 2004). It is common for school districts or programs to use lotteries to determine access to oversubscribed programs, and rigorous RCT evaluations have been conducted to eliminate the selection bias assumed to be associated with participation in voucher programs or charter schools, for example. A major limitation of education RCTs, however, is their external validity, as the samples of students subject to lotteries often are not representative of the broader population of students experiencing the educational treatment (e.g. Abdulkadiroglu et al., 2009). With few exceptions (e.g. Bifulco 2012), researchers are not able to determine how much selection bias is introduced when various quasi-experimental approaches are used in place of experimental ones. We are left wondering about the magnitude of the internal-for-external validity tradeoff that education researchers often face.

This study contributes to our understanding of this key methodological concern by assessing the extent to which quasi-experimental methods such as propensity score matching or observational models with control variables can replicate the “benchmark” experimental results of the original District of Columbia Opportunity Scholarship (DC OSP) evaluation conducted from 2004-2009 (see Wolf et al., 2010). The federal private school voucher program is an especially appropriate subject for such a methodological study because self-selection is assumed to be a major influence on whether or not a low-income urban student attends a private school.

We treat Instrumental Variables Analysis (IV) estimates of the impact of private schooling on student outcomes as the “benchmark” estimate of causal impact. A validated random lottery is the ideal instrumental variable with which to recover unbiased estimates of the effect of an intervention like private schooling in the face of substantial non-compliance with the original assignment of students to the treatment of private schooling through the mechanism of a voucher or the control condition. The lotteries used to create the experimental analysis sample for the DC OSP evaluation have been confirmed to have produced treatment and control groups with approximately similar baseline conditions (Wolf, Gutmann, Puma & Silverberg, 2006). Four years later, only 55% of the treatment group students were attending private schools while 11.5% of the control group “crossed-over” into private schools without a voucher. With the asset of a validated lottery and the problem of substantial experimental non-compliance, we argue that IV estimates are the most defensible benchmark to use in this case.

We compare the IV-generated benchmarks to the results from three types of alternative research designs for determining the effect of private schooling on student outcomes, in order (theoretically) from most- to least-biased: observational without controls (i.e. comparing simple...
group averages for private school students to public school students), observational with controls for baseline test scores and demographics, and propensity score matching. We interpret the extent to which the results from the alternative methods deviate from the results from the benchmarks as the degree of self-selection bias from employing that particular quasi-experimental method (e.g. Bifulco 2012).

We compare the results of these four methods across one attainment and two achievement outcomes: high school graduation rates, math achievement, and reading achievement, permitting us to test the robustness of our bias calculations to various outcomes.

We utilize student-level data contained in the restricted-use file associated with the original DC OSP evaluation. Student assessment data come from the Stanford Achievement Test-version 9 (SAT-9) administered by the DC OSP evaluation team. High school attainment data were based on parent surveys in which parents were asked if their child was still enrolled in high school, had graduated from high school, or fit neither category. The analytic samples include 1,328 student observations for the reading achievement outcomes, 1,330 for the math achievement outcomes, and 316 for the high school attainment outcomes.

The results of this study will be relevant to researchers and policy makers seeking to design or understand program evaluations in the future. If it appears that a quasi-experimental approach, when applied to the same set of students, yields effects estimates that are similar to those from an experimental benchmark, that would suggest that, under certain conditions, the use of a quasi-experimental approach might yield meaningful gains in external validity at a small cost to internal validity. If we find, on the other hand, that the methods (e.g. propensity score matching) generate results with large amounts of bias, such a finding would support the theory that RCTs are truly a “gold-standard” that cannot be easily approximated.

References


