Economics of Education Review 44 (2014) 100–113

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Horseshoes, hand grenades, and treatment effects? Reassessing whether nonexperimental estimators are biased

Kenneth Fortson*, Philip Gleason, Emma Kopa, Natalya Verbitsky-Savitz

Mathematica Policy Research, Human Services Division, 505 14th Street, Suite 800, Oakland, CA 94612, USA

ARTICLE INFO

Article history:
Received 20 March 2013
Revised 27 October 2014
Accepted 2 November 2014
Available online 14 November 2014

JEL Codes:
I21
C10
C18
C21.

Keywords:
Treatment effects
Randomized controlled trials
Nonexperimental methods
Within-study comparison

ABSTRACT

Randomized controlled trials (RCTs) are considered the gold standard in estimating treatment effects. When an RCT is infeasible, regression modeling or statistical matching are often used instead. Nonexperimental methods such as these could produce unbiased estimates if the underlying assumptions hold, but those assumptions are usually not testable. Most prior studies testing nonexperimental designs find that they fail to produce unbiased estimates, but these studies have examined weaker evaluation designs. The present study addresses these limitations using student-level data based on a large-scale RCT of charter schools for which standardized achievement tests are the key outcome measure. The use of baseline data that are strongly predictive of the key outcome measures considerably reduces but might not completely eliminate bias.

1. Introduction

Experimental evaluations based on randomized controlled trials (RCTs) are widely considered to be the gold standard in evaluating the effects of a social program. However, an RCT is not always feasible. In some contexts, it might not be logistically possible or ethical to exclude individuals from participating in the program. In other contexts, researchers seeking to estimate the treatment effect of a program might lack the authority or resources to employ a random assignment design, even if it were logistically possible. Even when random assignment is possible for an intervention, it might not be possible for everyone served by the intervention, in which case the findings might not generalize broadly. For example, the experimental analysis of charter schools by Gleason, Clark, Tuttle, and Dwoyer (2010), on which the current study is based, used lotteries employed by oversubscribed charter schools. Though their evaluation design had strong internal validity, the findings do not generalize to charter schools that were not oversubscribed.

When an RCT is infeasible, researchers often resort to a nonexperimental approach for estimating program effects. A popular class of nonexperimental designs uses a nonrandomly selected comparison group to represent what would have happened to the treatment group had they not participated in the program. However, the assumptions underlying nonexperimental evaluations are usually not testable in practice. This study examines the validity of comparison group designs based on regression and propensity score matching (PSM) using data from an experimental evaluation of charter schools (Gleason et al., 2010), testing whether these designs can replicate the findings from a well-implemented random assignment study.

In an experimental evaluation design, the randomly assigned control group is used to estimate the counterfactual—what would have happened in the absence of the

* Corresponding author. Tel.: +1 5108303711.
E-mail address: kfortson@mathematica-mpr.com (K. Fortson).

http://dx.doi.org/10.1016/j.econedurev.2014.11.001
0272-7757/

intervention. When implemented well, an RCT ensures that the control group does not differ from the treatment group in any systematic way that could bias the estimated treatment effect. In contrast, a comparison group design estimates the counterfactual using a group that was not exposed to the intervention for any number of nonrandom reasons. Comparison group methods can, in theory, produce estimated treatment effects that are as good as those of a well-implemented experimental design. However, even the best comparison group designs rely on the assumption that the analysis can adjust for any differences between the characteristics of the treatment and comparison groups prior to treatment, and that on average, the two groups do not differ on any other unobserved dimensions that are correlated with the outcome(s) of interest (Rosenbaum & Rubin, 1983; Little & Rubin, 2000).

One approach to investigating the question of whether comparison group methods produce unbiased treatment effect estimates involves efforts to replicate estimates from an existing experimental study using a comparison group design—a validation approach that is referred to in the literature as a "replication study" or a "within-study comparison." A within-study comparison starts with a well-implemented experimental study that can be credibly believed to have produced unbiased estimates and then applies a comparison group design to estimate the same treatment effect parameters using data collected at least in part in the same study.

Most of the existing replication studies of comparison group designs have been conducted for evaluations of job training programs, and the majority of these have found that comparison group designs cannot reliably replicate experimental estimates. This was the conclusion of the early replication work of Lalonde (1986), Fraker and Maynard (1987), and Friedlander and Robins (1995), and has been a consistent finding in most subsequent replication studies, as summarized by Glazerman, Levy, and Myers (2003). An exception was the work by Dehejia and Wahba (1999), which found that PSM methods could replicate experimental results. Smith and Todd (2005) subsequently found that these results were not robust to minor changes in the analysis sample, though Dehejia and Wahba dispute some of the Smith and Todd findings in further correspondence between the two sets of authors. Dehejia and Wahba's findings were also sensitive to the pre-intervention variables used, suggesting that rich pre-intervention data are necessary to overcome possible selection on observables. Recent work by Bloom, Michalopoulos, and Hill (2005) and Peikes, Moreno, and Orzol (2008) has expanded replication studies to other contexts, but the basic findings have been the same.

Education interventions are attractive for a within-study comparison because achievement test scores are often the outcomes of greatest interest. Because achievement test scores are highly correlated over time, baseline measures of this outcome are likely to be highly predictive of follow-up measures of the outcome. Achievement test scores are also measured uniformly for most students in the same grade, at least within a locality and often within an entire state. Despite these advantages, few within-study comparisons have attempted to replicate experimental estimates of educational interventions' treatment effects. Two early exceptions are the within-study comparisons by Agodini and Dynarski (2004) and Wilde and Hollister (2007), which base their analyses on a drop-out prevention program and the Tennessee Project Star class size experiment, respectively. Both studies conclude that nonexperimental methods fail to replicate experimental findings. However, neither study was able to control for pre-intervention measures of the outcome. More recently, Bifulco (2012) examined magnet schools near Hartford, Connecticut and found that propensity score methods could come close to replicating the experimental findings when highly predictive baseline data were used. Abdulkadiroglu, Angrist, Dynarski, Kane, and Pathak (2011) consider whether regression models come close to replicating experimental findings as part of a broader study of Boston's charter and pilot schools, though the within-study comparison is not the focus of their study, and consequently, their comparison of experimental and nonexperimental estimates is less formal and inconclusive.

Cook, Shadish, and Wong (2008) and Shadish, Clark, and Steiner (2008) argued that the failure of comparison group designs to replicate experimental results stems from differences in data sources or unsuitable comparison groups. Cook et al. (2008) describe conditions that efforts to validate nonexperimental methods via a within-study comparison with a randomized experiment should attempt to meet. Key among them are that the experimental and nonexperimental approaches must be demonstrably good examples of their types, and the data sources should be the same for the two analyses. The analyses should estimate the same statistical relationship. For example, if the experimental benchmark is an estimated effect of the intent to treat (ITT), the nonexperimental estimates should estimate the ITT effect, too.

The within-study comparison presented in this paper contributes to the existing body of knowledge in two main ways. This study is one of the few replication studies of comparison group designs that (1) focuses on an education intervention and outcomes allowing us to control for pre-intervention measures of the outcome and (2) examines nonexperimental designs using a within-study comparison approach that addresses the concerns described in Cook et al. (2008) and Shadish et al. (2008). In contrast to previous work, key features of the present study are that our comparison group is drawn from same local areas as the experimental sample; we applied each approach such that the target parameter we are estimating is the same; and we systematically compare the two sets of estimates based on objective criteria, in contrast to previous studies that have only done subjective, ad hoc comparisons. Our study also has the advantage that, rather than being limited to one city, it uses data from 15 localities across six states. Consequently, any idiosyncrasies in one or two sites are less likely to determine whether our nonexperimental analyses replicate the experimental findings.

The nonexperimental approaches that we examine in this paper were specified in a research protocol in advance of the analysis, as were the two criteria we use to determine whether a given nonexperimental treatment effect estimate replicates the experimental benchmark. The first criterion for assessing equivalence is to consider whether the conclusion that would be drawn from its estimated treatment effect is the same. Specifically, we examine whether the basic magnitude and sign of the estimates are comparable and whether the statistical significance (or insignificance) is the same. The second criterion is whether the nonexperimental estimate is statistically different from the experimental benchmark.