Journal of Economic Behavior & Organization xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

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

The performance of non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark

Paul J. Ferraro [a,*], Juan José Miranda [b]

[a] Department of Economics, Andrew Young School of Policy Studies, Georgia State University, PO Box 3992, Atlanta, GA 30302-3992, United States
[b] Economics Unit, Sustainable Development Department, The World Bank, 1818 H St., NW, Washington, DC 20433, United States

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form 28 February 2014
Accepted 15 March 2014
Available online xxx

Keywords:
Causal inference
Impact evaluation
Nonpecuniary
Within-study
Selection-on-observables

ABSTRACT

In the field of environmental policy, randomized evaluation designs are rare. Thus researchers typically rely on observational designs to evaluate program impacts. To assess the ability of observational designs to replicate the results of experimental designs, researchers use design-replication studies. In our design-replication study, we use data from a large-scale, randomized field experiment that tested the effectiveness of norm-based messages designed to induce voluntary reductions in water use. We attempt to replicate the experimental results using a nonrandomized comparison group and statistical techniques to eliminate or mitigate observable and unobservable sources of bias. In a companion study, Ferraro and Miranda (2013a) replicate the experimental estimates by following best practices to select a non-experimental control group, by using a rich data set on observable characteristics that includes repeated pre- and post-treatment outcome measures, and by combining panel data methods and matching designs. We assess whether non-experimental designs continue to replicate the experimental benchmark when the data are far less rich, as is often the case in environmental policy evaluation. Trimming and inverse probability weighting and simple difference-in-differences designs perform poorly. Pre-processing the data by matching and then estimating the treatment effect with ordinary least squares (OLS) regression performs best, but a bootstrapping exercise suggests the performance can be sensitive to the sample (yet far less sensitive than OLS without pre-processing).

1. Introduction

Environmental programs are rarely implemented within a randomized design. To estimate program impacts, researchers must therefore depend on non-experimental, observational designs. Although observational designs have experienced important advances over the last two decades, debate continues about how well these designs can uncover causal

* Corresponding author. Tel.: +1 404 413 0201.
E-mail addresses: pferraro@gsu.edu (P.J. Ferraro), jjmiranda@worldbank.org (J.J. Miranda).

http://dx.doi.org/10.1016/j.jebo.2014.03.008
0167-2681/

relationships (Smith and Todd, 2005). To infer causality from observational designs, one must rely on untestable assumptions about differences in the potential outcomes between treatment and control groups.

Starting with Lalonde (1986), a small but growing number of social scientists have tried to use randomized experiments to validate observational evaluation designs. In these studies, called "within-study comparisons" or "design replication studies," researchers first estimate a program's impact by using a randomized control group. Under effective randomization, the estimated impact is assumed to have high internal validity. Then the researchers re-estimate the impact by using one or more nonrandomized comparison groups and statistical techniques to eliminate or mitigate observable and unobservable sources of bias. If the non-experimental estimate is close to the experimental estimate, the non-experimental design is labeled "successful" (how to define "close" is explored below).

Lalonde (1986), Fraker and Maynard (1987) and Lalonde and Maynard (1987) use data from the National Supported Work randomized field experiment and form non-random comparison groups using data from national surveys. The authors argue that non-experimental designs cannot systematically recover the experimental estimates. Using the same data, but a different empirical design, Dehejia and Wahba (1999, 2002) come to the opposite conclusion, but Smith and Todd (2005) question the robustness of their results. Uncertainty about the ability of non-experimental designs to systematically replicate the results of experimental designs is also reflected in design replication studies using data from educational programs (Agodini and Dynarski, 2004; Hill et al., 2004; Wilde and Hollister, 2007), poverty reduction programs (Diaz and Handa, 2006; Handa and Maluccio, 2010), migration (McKenzie et al., 2010), and elections (Arceneaux et al., 2006).

Our design-replication advances the literature in five important ways. First, with the exception of a companion study (Ferraro and Miranda, 2013a), we know of no design-replication study in environmental policy and none that focus on information-based interventions. Second, our study context is one in which exposure to the program occurs because of where a person chooses to live rather than self-selection. Like in other policy contexts, environmental policies and programs are frequently implemented or piloted in administrative units like towns, counties, or states. To estimate impacts, scientists typically look to neighboring administrative units to form comparison groups and apply various statistical techniques to control for observable and unobservable sources of bias. Such contexts and research approaches are not, of course, specific to environmental policy; see, for example, the seminal article by Card and Krueger, 1994, who use Pennsylvania fast-food restaurants as controls for New Jersey fast-food restaurants in order to assess the impact of minimum wage laws (other noteworthy examples include Besley and Burgess, 2004; Galiani et al., 2005). We know of no design-replication studies in such a policy context.

Third, our study is the first to contrast how the performance of an observational design is affected when one moves from a rich data environment, with many observable covariates and repeated outcome observations before and after treatment assignment, to a less rich data environment. Panel data designs with repeated pre- and post-treatment observations are uncommon in the environmental evaluation literature. Understanding how standard statistical techniques perform without such rich panel data is therefore important.

Fourth, we study two treatments: one that had a large, statistically significant impact and another that had a negligible, statistically insignificant impact. To our knowledge, the design-replication literature comprises only contexts in which there is a single treatment that is known to have had a statistically significant, policy-relevant impact. A valid observational design should be able to detect an impact where one exists and fail to detect one where one does not exist.

Fifth, our study includes observational designs that are rare or absent in the design-replication literature. Most design replication studies focus on the performance of propensity score matching (PSM). In addition to applying matching designs, we also apply inverse probability weighting (IPW) designs, as well as combinations of matching, trimming, IPW and ordinary least squares (OLS) regressions.

Our experimental benchmark comes from a 2007 randomized field experiment. The experiment tested the effects of conservation messages on voluntary reductions in water use among over 100,000 households during a drought in metropolitan Atlanta, Georgia, USA (Ferraro and Price, 2013; Ferraro and Miranda, 2011, 2013b). Assuming that there were no randomization biases or general equilibrium (spillover) effects, which were not detected in Ferraro and Price (2013), randomization of households into control and treatment groups, followed by a comparison of each group's mean water consumption, provides an unbiased estimator of the average treatment effect (ATE).

To form a **non-experimental** comparison group, we use households from a neighboring county. The neighboring county had similar water pricing policies and the same water sources, weather patterns, state and metro regulatory environments and other regional confounding factors during the messaging experiment. Participants did not self-select into the program, but they may have sorted themselves across counties based on characteristics that also affect water consumption. By merging water use data with tax assessor and census data, we create a unique data set that includes measures of household water use, home characteristics, and block group characteristics.

Our context satisfies five of the six criteria proposed by Cook et al. (2008) for high-quality design-replication studies (like all published design-replication studies, our study fails the fifth criterion, which recommends that analysts of the observational data be blind to the results of the randomized experiment). Our context also satisfies three criteria that Heckman and colleagues have argued are needed to draw unbiased (or small bias) inferences from observational designs (Heckman et al., 1997, 1998a, 1998b): (i) participant and nonparticipant data come from the same sources, with similar measures of the outcome variable being most important; (ii) participants and nonparticipants share the same economic environment; and (iii) the data contain a rich set of variables that affect both program participation and the outcome.

Please cite this article in press as: Ferraro, P.J., Miranda, J.J., The performance of non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.03.008