European Journal of Political Economy xxx (2017) 1–8

Contents lists available at ScienceDirect

European Journal of Political Economy

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

Randomized controlled trials informing public policy: Lessons from project STAR and class size reduction

Moshe Justman [a, b, *]

[a] School of Economics and Business Administration, Ruppin Academic Center, Israel
[b] Department of Economics, Ben Gurion University of the Negev, Israel

ARTICLE INFO

JEL classification:
C54
I28

Keywords:
Class size
Project STAR
Randomized controlled trials
Field experiments
Validity
Modern experimentalist paradigm

ABSTRACT

Randomized controlled trials (RCTs) and related research strategies are increasingly seen as the preferred methodology for evaluating policy interventions, but a single-minded focus on identifying causal effects limits their capacity to support actual policy decisions. A detailed look at Project STAR illustrates this point and offers some possible correctives. Initiated by the Tennessee legislature to help decide whether to enact state-wide class size reduction (CSR), STAR compared the test scores of students randomly assigned to classes of different size. It addressed a well-formed research question, but focused narrowly on refining a single link in a long chain of evidence necessary to address the policy question at hand: whether CSR would be a good use of a large increase in education spending. It disregarded the limitations of test scores as indicators of education quality and ignored general equilibrium effects on teacher quality and salaries. Moreover, the emphasis it placed on estimating average CSR effects in a given setting diverted attention from the heterogeneity of these effects and the conditions that mediated their impact, limiting its external validity. These observations continue to be relevant for the design of policy-oriented research, and for the academic training of empirical economists.

1. Introduction

The use of randomized controlled trials (RCTs) and other forms of randomized assignment to establish causality is a central element of what Angrist and Pischke (2009) call the "modern experimentalist paradigm" (MEP). Policy-oriented empirical studies are often susceptible to selection biases that can distort findings, and randomization has become the "Gold Standard" for eliminating these biases. The Institute of Economic Studies' What Works Clearinghouse (WWC) sees randomization as a necessary condition for classifying a study as one that "meets WWC standards without reservation."[1] Similarly, MIT's highly influential Abdul Latif Jameel Poverty Action Lab (J-PAL) for policy-oriented research on developing economies aims "to support the use of randomized evaluations … and to encourage policy changes based on results of randomized evaluations."[2] Indeed, as Deaton and Cartwright (2016) note, there are many other such "… 'What Works' centers using and recommending RCTs in a huge range of areas of social concern across Europe and the Anglophone world … "

The MEP as commonly practiced follows Rubin's (2005) call to "separate scientific inference for causal effects from decisions based

* School of Economics and Business Administration, Ruppin Academic Center, Israel
E-mail address: justman@bgu.ac.il.

[1] http://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_info_rates_061015.pdf.

[2] https://www.povertyactionlab.org/about-j-pal See also Duflo's (2017) Ely Lecture, which elaborates on the important contribution of this approach to getting right the details of policy planning and implementation.

http://dx.doi.org/10.1016/j.ejpoleco.2018.04.005
Received 9 November 2017; Received in revised form 1 April 2018; Accepted 16 April 2018
Available online xxxx

on such inference," focusing its efforts on questions of economic or social causality for which precise, unbiased answers can be obtained without a need for theory. This approach has produced persuasive findings on a wide range of empirical issues but its single-minded focus on "theory-free" identification of causal effects may inhibit its capacity to provide a useful evidence base for deciding concrete policy issues.[3]

The present paper argues this point through a detailed look at Project STAR, a rigorous and extensively analyzed, large-scale RCT commissioned by the Tennessee legislature in 1985 to help it decide whether to mandate statewide class-size reduction (CSR) in kindergarten to grade three (K-3) from 22 students per class to 15 students. Its careful research design sought to determine whether *ceteris paribus* reducing class size improves test scores, by randomly assigning K-3 students to classes of either (approximately) 22 or 15 students, and comparing tests scores across class size.

The policy question Project STAR was meant to address, was whether CSR would be a good use of a large increase in Tennessee's education spending, worth the tax increase necessary to fund it, and a better use of these funds than alternative school policy reforms.[4] Instead it focused its considerable efforts on answering a well-defined but much narrower research question—identifying the causal effect of class size on student test scores.[5] Thus it ignored the potentially wider benefits of smaller classes for early child development that underlie much of the broad political support for CSR in early grade levels (Schrag, 2006). It ignored general equilibrium effects of a surge in demand for early childhood teachers on teacher quality and salaries (Jepsen and Rivkin, 2009), and the impact of smaller classes in public education on private school enrolment (Gilraine et al., 2018); and it ignored the negative externality of a general rise in test scores on the value of individual scores.[6] Without addressing these essential elements, identifying causal effects more or less precisely can have little practical value for policy decisions.

In addition, the emphasis STAR placed on estimating average CSR effects in a non-random sample of Tennessee schools diverted attention from the heterogeneity of these effects, while its theory-free approach provided no indication of conditions under which CSR is more or less effective. This limited the external validity of its findings for other, different settings, illustrating Cartwright and Munro's (2010) broader argument, that asking whether a proposed social policy has a significant effect on desired outcomes is often less useful than understanding the factors that determine its capacity to achieve such an effect.[7]

These limitations of STAR were largely overlooked in published analyses of its findings, which were largely interpreted as unequivocally supporting CSR, and which played a part in the subsequent spread of CSR policies across the United States. This may well have been a "good thing", as the exclusive focus of STAR on test scores clearly underestimated its benefits, but a more eclectic approach might have anticipated some of the problems that arose in the implementation of CSR, and generated more-realistic expectations. In the event, large-scale expensive CSR initiatives such as those undertaken in California and Florida, encountered unanticipated problems, and subsequent assessments of their impact on test scores fell short of expectations.

The present study speaks to two strands of the literature. One comprises broadly framed critiques of the paradigmatic status of RCTs in economics by Heckman (2000), Heckman and Vytlacil (2007), Deaton (2010), Sims (2010), Basu (2013), Hausmann (2016), Rothstein and von Wachter (2016) and Deaton and Cartwright (2016) among others.[8] The detailed look offered here at a concrete example highlights many of the important points raised in these more general treatments, while indicating how such shortcomings might be addressed in practice. A second relevant strand of the literature comprises internal critiques of RCTs, mostly in development economics, which increasingly recognize many of these challenges to external validity, often referred to as the challenge of scaling-up. Prominent examples are Al-Ubaydki et al. (2017), Banerjee et al. (2017), Muralidharan and Niehaus (2017), Banerjee et al. (2016) and Gilraine et al. (2018).[9] The present paper uses the example of Project STAR as an organizing device that provides an integrated perspective on factors that limit the scope for RCTs shaping public policy, in a different context, and how they might be addressed.

The shortcomings of Project STAR are typical of social policy RCTs, which generally suffer from weak external validity because they tend to focus on average effects, and lack a theoretical framework to explain heterogeneity; do not adequately address general equilibrium effects; and often focus on a well-defined research question that fails to capture politically relevant dimensions of the issue at hand. They are typical of RCTs because they reflect the excesses of the MEP, which underpins much of the current published evaluations

[3] See also Athey and Imbens' (2017) related discussion. This limitation is distinct from the inhibiting effect of stringent data requirements, implicit in the MEP, on the universe of questions that can be addressed.

[4] Our main focus is on Project STAR as designed. A closer look at the implementation of STAR in the following section reveals significant departures of implementation from design, some of which seem unavoidable in a social-policy setting. Similar departures are documented by Ginsburg and Smith (2016) in their analysis of 27 RCTs on mathematics curricula, and by Necker (2014) in her survey of economists' "misbehavior". These are not our primary focus here. Some RCTs raise serious ethical concerns (Greenberg et al., 1999). These are important, but again not our primary focus here.

[5] Identifying optimal class size is the first in a list of research questions that Angrist and Pischke (2009) recommend to their readers as subjects for empirical research within the MEP. Equating better learning with improved scores on a narrow set of tests that capture only a small part of learning, and even less of the other important things schools do, is a fundamental weakness of this approach.

[6] Rubin (2005) calls such effects departures from the stable unit treatment value assumption (SUTVA). As Imbens (2009) notes, questions involving such effects cannot be answered by simple experiments.

[7] In this spirit, Banerjee et al. (2017, p. 80) responding to Pritchett and Sandefur's (2015) emphasis on context dependence, recognize that "once we admit the need for a prior for aggregating results, there is no reason to stick to purely statistical approaches. An alternative is to use the existing evidence to build a theory, which tries to account for why some experiments succeed and others fail." They cite Kremer and Glennerster's (2011) work on the take-up of preventive health products as an example.

[8] Parallel critiques of RCTs in general scientific contexts include Vandenbroucke (2004), Worrall (2007), Cartwright (2007), Ullman (2015), and Vandenbroucke et al. (2016). These critiques notwithstanding, the example of medical research greatly advanced the adoption of RCTs in economics; when the World Bank announced its RCT program, The Lancet (2004) rejoiced "The World Bank is finally embracing science" (Deaton, 2010).

[9] Less directly, the present analysis complements Aron-Dine, Einav and Finkelstein's (2013) meticulous re-examination of the 1974 RAND Health Insurance Experiment, which focused on departures of its implementation from an ideal experimental design.