RESEARCH SUPPORTING POLICY, A HISTORICAL PERSPECTIVE

Moshe Justman

Discussion Paper No. 20-14

August 2020

Monaster Center for
Economic Research
Ben-Gurion University of the Negev
P.O. Box 653
Beer Sheva, Israel

Fax: 972-8-6472941
Tel: 972-8-6472286

Research Supporting Policy, A Historical Perspective[1]

Moshe Justman[2]

The Covid-19 pandemic highlights fundamental questions about the relation between academic expertise and policy formation. Epidemiologists, who have the most relevant scientific expertise for dealing with pandemics, seem the least sure of their recommendations. Aware of the many unknowns, they respond with caution to an amorphous situation, drawing on a methodology that combines theoretical hypotheses with weakly discerned parameters, far removed from the medical paradigm of randomized controlled trials (RCTs) that produce definitive answers to clear-cut questions.

Academic, policy-oriented, economic research—on health, education, welfare, and related issues—used to be more like epidemiological research, drawing heavily on mathematical models calibrated to "best-guess" parameters, but has gravitated more and more toward the medical paradigm. Increasingly, policy-oriented economic research invests its energies in producing unassailable, causal evidence—through RCTs and other randomization strategies—on narrowly defined empirical hypotheses, preferably without the distraction of theory. These elaborate, expensive studies are strong on internal validity, but seem lost in the dynamic, wildly-changing, policy environment that the Covid-19 crisis has created.

The rise of the RCT in the economics of education

The arc that the economics of education has followed over the last half century nicely illustrates this paradigmatic change in policy-oriented economic research. These are relatively recent developments. The very idea that education is a suitable object of economic analysis came into its own, as the theory of human capital, only in the 1960s with the seminal contributions of Theodore Schultz, Gary Becker, and their students at the University of Chicago.[3] They gave mathematical form to the metaphor of education as investment in a child's future, and *took it to the data*, demonstrating a robust statistical association between people's schooling and work experience, and their labor income. This cast wages as the return on a stock of human capital, accumulated through investment in schooling. It indicated a path to a systematic weighing of the (easily measured) costs of schooling against what now seemed to be its quantifiable benefits, and the prospect of a new *evidence-based* approach to education policy.

However, this approach clearly needed more work; these statistical associations could not be construed as measures of causal effects. One obvious problem was "self-selection." The average difference in earnings between, say, college and high school graduates cannot be attributed entirely

[1] Justman, Moshe. "Research Supporting Policy, A Historical Perspective." Series, "Policy Models in Pandemic." *Items: Insights from the Social Sciences*, July 30, 2020. https://items.ssrc.org/covid-19-and-the-social-sciences/policy-models-in-pandemic/research-supporting-policy-a-historical-perspective/

[2] Department of Economics, Ben Gurion University, Beer-Sheva, Israel. justman@bgu.ac.il

[3] For millennia, the purpose of education was defined in religious terms, in terms of social responsibility, and later, of personal development, as in Jean-Jacques Rousseau's highly influential *Emile, or On Education*.

to the benefits of a college education, as those who choose to attend college generally expect to gain more from their education.[4] Various remedies were proposed—among them James Heckman's structural econometric approach—but these were overtaken by the "gold standard" of current econometric practice: randomization, "natural experiments" and RCTs that produce statistically significant, unbiased estimates of causal effects without the need for debatable theoretical constructs.

While RCTs effectively address the problem of self-selection, they exacerbate a second class of problems, challenges to external validity. Research conducted under controlled, experimental conditions is inherently subject to close supervision, and heightens subjects' self-awareness, compared to the conditions of routine implementation. Specific social, cultural, or institutional conditions, in which an experiment takes place, may undermine its validity for other contexts. Moreover, experimental results may not scale up well, often because RCTs are explicitly designed to neutralize general-equilibrium effects, which later prove crucial in full-scale implementation. In some cases, there is a disconnect between the research question on which the RCT focuses and the policy issue on which it was meant to shed light.

The instructive example of Project STAR

Project STAR, among the best known and most extensively analyzed, large-scale education RCTs, illustrates all of these points. The Tennessee legislature commissioned the study in 1985, at considerable cost, to help decide whether to enact a statewide reduction of class size in kindergarten to grade three (K–3), from 22 to 15 students. Its research design focused on whether such class-size reductions (CSRs) improve tests scores, in themselves, holding other factors constant. To this purpose, the study randomly assigned a sample of a few thousand kindergarten students to classes of either (approximately) 22 or 15 students, and compared their test scores each year, to determine whether, within each school, students in the smaller classes had significantly higher scores.[5]

There were several weaknesses in this research design as it related to the original purpose of the legislators who commissioned it. Reducing class size from 22 to 15 students increases the number of classes by about 50 percent, entailing a budget increase of roughly similar magnitude. The relevant policy question is presumably not whether this sizable budget increase had *any* statistically significant, causal effect on K–3 test scores, but whether the gain could justify the extra cost; or at the very least, whether CSR represented a better use of these extra resources than, say, a commensurate increase in teacher salaries.[6]

[4] Of course, other factors also influence college enrollment, among them liquidity and social mores, which may offset differences in expected gains; and measurement error in right-hand variables reduces estimated gains.

[5] This is a schematic description. A closer look at its implementation reveals various actual departures from its designed randomization. For further details and references, see Moshe Justman, "Randomized Controlled Trials Informing Public Policy: Lessons from Project STAR and Class Size Reduction," *European Journal of Political Economy* 54 (Sept. 2018): 167–174.

[6] Most analyses found an overall improvement of about 0.2 standard deviations in test scores, most of it in kindergarten, see Alan B. Krueger, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114, no. 2 (1999): 497–532. Efforts to quantify the long-term impact of this gain on lifetime income suffered from various methodological weaknesses. They incorporated estimates of the

A further limitation of the study was its failure to acknowledge the general equilibrium impact of increased teacher demand on the level and distribution of teacher quality. This had substantial real-world implications. The positive effect on test scores it found helped promote similar initiatives in other states, among these California, where the pressing need for more teachers led high-income school districts to hire experienced teachers away from low-income districts. The low-income districts then had to fill their ranks with inexperienced—sometimes unaccredited—teachers, resulting in a widening of test-score gaps between high and low-income districts.

Yet, a third weakness of Project STAR was its exclusive focus on test scores. The advantage of smaller classes in grades K–3 for parents and educators alike is not only the prospect of a small average increase in test scores—statistically significant and causal, but still only 0.2 standard deviations. Smaller classes also allow teachers to get to know their students better. In the early grades, this allows teachers a better opportunity to identify previously undiagnosed health and developmental problems, difficult situations in the home, learning disabilities, and so on.

The exclusive focus of Project STAR on test scores mirrors a generally disproportionate focus of economists' policy-oriented education research on test results and, more generally, on quantifiable outcomes. In some respects, this has worsened over time. Early analyses of the benefits of education in the human capital framework focused on its contribution to wages and income, and later on its wider benefits for life choices—smoking, drug addiction, teen-age pregnancies. These studies pose challenges to establishing causality. Consequently, research on the effectiveness of teachers and schools has come to focus more and more on standardized test scores in a narrow range of subjects as its measure of outcomes.

This is an outgrowth of the impossibly high standards of internal validity, which have come to determine what counts as acceptable evidence.[7] Outcomes that are difficult to quantify, and outcomes that are difficult to assign to specific causes, do not fit this mold. This severely limits the usefulness of leading-edge research in the economics of education, much of it devoted to establishing to the highest methodological standards findings that most people think they knew all along—such as that schoolchildren do better in smaller classes. Economists will say, "You didn't actually *know* this until I established causality." But then most people will answer, "Yes, I did."

Will the Covid-19 crisis leave a lasting mark on policy-oriented research?

The Covid-19 crisis has improved the image of useful but imperfect evidence in support of public policy. The natural experiments that health economists continue to seek in the heat of the pandemic still offer prospects of prestigious publication, but shed little light on the pressing problems at hand; while epidemiologists estimating highly sensitive, model-dependent reproduction coefficients from biased statistics on daily infection rates are able to offer useful practical guides to policy.

relation between test scores and income without causal validity; equated the benefit of higher test scores for individuals with the social benefit of higher scores; and ignored departures from randomization in actual implementation; see Justman, "Randomized Controlled Trials Informing Public Policy."

[7] Ironically, these are more stringent than the criteria applied in criminal courts, where all that is required is proof beyond reasonable doubt.

Respected research institutions, eager to be a part of this collective effort, now regularly issue working papers that suffer from severe methodological flaws but are arrestingly interesting. Some are even finding their way to well-regarded journals. Is this the beginning of a new, sustained trend? Will we see more imperfect but interesting papers crowding out unassailable RCT proofs and robustness checks of facts we thought we knew? Or will the winds of change subside when the crisis is over?