THE UNIVERSITY OF MELBOURNE

FACULTY OF BUSINESS & ECONOMICS

Melbourne Institute Working Paper Series

Working Paper No. 37/16

Economic Research and Education Policy: Project STAR and Class Size Reduction

Moshe Justman

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Economic Research and Education Policy: Project STAR and Class Size Reduction*

Moshe Justman
Melbourne Institute of Applied Economic and Social Research, The University of Melbourne; and Department of Economics, Ben Gurion University of the Negev

Melbourne Institute Working Paper No. 37/16

ISSN 1447-5863 (Online)

ISBN 978-0-73-405234-6

December 2016

* This is an expanded and revised version of my Patinkin Lecture, at the Annual Meetings of the Israel Economic Society, May 2016. Sincere thanks to Shmuel Nitzan for inviting me to give the lecture and to Arye Hillman for encouraging me to write this English version. I am also grateful for their comments to Nachum Blass, Henry Braun, Danny Cohen-Zada, Avital Darmon, David Figlio, Naomi Friedman-Sokuler, Mark Gradstein, Gadi Hazak, Guyonne Kalb, Shirlee Lichtman-Sadot, Cain Polidano, Dave Ribar, Oren Rigbi, Jesse Rothstein, Nico Salamanca, Michel Strawczynski, Ro'i Zultan, and seminar participants at the Melbourne Institute of Applied Economic and Social Research. None bears any responsibility for any of the content of this paper. For correspondence, email <justman@bgu.ac.il>.

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne
Victoria 3010 Australia
Telephone **(03) 8344 2100**
Fax **(03) 8344 2111**
Email **melb-inst@unimelb.edu.au**
WWW Address **http://www.melbourneinstitute.com**

Abstract

The use of randomized controlled trials (RCTs) and related randomization strategies to eliminate selection biases in establishing causality is a key element of the "modern experimentalist paradigm" (MEP). Yet, its emphasis on precisely identifying causal factors often limits its capacity to provide an evidence base for policy. We illustrate this through a detailed look at Project STAR, an extensively analyzed, well-funded, large-scale, rigorous RCT commissioned by the Tennessee legislature to help it decide whether to mandate statewide class-size reductions (CSR) from kindergarten to the third grade. Project STAR randomly assigned students to classes of different size and compared test results across these classes, to obtain an unbiased answer to the research question, "Does reducing class size improve test scores?" However, this shed little light on whether reducing class size was a good use of increased education financing. Analyses of Project STAR ignored general equilibrium effects of CSR on both the demand for teachers and the value of test scores. Moreover, its emphasis on estimating average class-size effects in a particular setting diverted attention from their heterogeneity, and the need to understand how class size affects learning, and how its effect is moderated by circumstances. Rather than considering the full chain of evidence necessary for shaping class-size policy, Project STAR concentrated its effort on maximizing the accuracy of a single link in that chain; internal validity trumped policy relevance.

JEL classification: C54, I28

Keywords: Class size, Project STAR, randomized controlled trials, field experiments, internal validity, external validity, modern experimentalist paradigm

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. Thus, the United States Department of Education Institute of Economic Studies What Works Clearinghouse (WWC) sees randomization as a necessary condition for classifying a study as "meets WWC standards without reservation."[1] Similarly, the stated objective of MIT's highly influential Abdul Latif Jameel Poverty Action Lab (J-PAL) for policy-oriented research on developing economies, is "to support the use of randomized evaluations, to train others in rigorous scientific evaluation methods, and to encourage policy changes based on results of randomized evaluations."[2] And 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 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 issues but is less useful as a guide to policy decisions. The same methodological decisions that reduce bias and promote greater accuracy in its findings also undermine its capacity to provide a useful evidence base for specific policy issues, even as its stringent data requirements severely restrict the universe of questions it is able to address. Where there is a tradeoff between policy relevance and internal validity, the MEP favors internal validity.

We illustrate this difficulty through a detailed look at Project STAR, a rigorously executed and extensively analyzed, well-funded, large-scale RCT. Commissioned by the Tennessee legislature in 1985 to help it decide whether to mandate statewide class-size reductions (CSRs), Project STAR randomly assigned students from kindergarten to grade three (K-3) to classes of either (approximately) 22 or 15 students, and compared tests scores across different size classes to determine whether reducing class size improves test scores. A close look at its

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

[2] https://www.povertyactionlab.org/about-j-pal

design and implementation reveals various departures of implementation from design.[3] However, these are tangential to our central argument, which is that Project STAR *as designed* could not provide a useful guide to shaping the policies it was meant to inform. These limitations were largely overlooked in published analyses of Project STAR. Consequently, in the subsequent spread of CSR policies across the United States, problems that might have been anticipated came as a surprise, and the perceived benefits of expensive, large-scale CSR initiatives such as those in California and Florida fell short of expectations.

This detailed study of Project STAR complements more broadly framed critiques of the paradigmatic status of RCTs in economics by Heckman (2000), Heckman and Vytlacil (2010), Deaton (2010), Sims (2010), Basu (2013), Hausmann (2016), Rothstein and von Wachter (2016) and Deaton and Cartwright (2016) among others.[4] The careful look it offers at a well-funded, highly regarded, extensively analyzed RCT adds concreteness to many of the important points raised in these more general treatments, and shows that the potential flaws which they indicate can arise even under seemingly ideal conditions. It also complements Aron-Dine, Einav and Finkelstein's (2013) meticulous re-examination of the RAND Health Insurance Experiment, initiated in 1974 at a cost of 300 million dollars (at current prices), and still widely viewed as the "gold standard" in its field. Where their re-analysis of the RAND data focused on departures of implementation from an ideal experimental design,[5] our central argument is that Project STAR as designed could not provide a useful guide to shaping the policies it was meant to inform.

This stems, in the first instance, from the research question it addressed, "Does a reduction of class size increase learning?", a well-formulated research question tailored to obtain a clear, unbiased answer through randomized experimentation.[6] But it is not the policy question that

[3] Such departures are largely inevitable in social policy research. For related work, see, e.g., Ginsburg and Smith's (2016) analysis of 27 RCTs on mathematics curricula in the What Works Clearinghouse, and Necker (2014), on a survey of economists' divergences from accepted practice. A broader discussion would also address ethical dimension of social experiments (Greenberg, Shroder and Onstott, 1999).

[4] In this, economics follows evidence-based medicine. As *The Lancet* (2004) announced when the World Bank announced its program of RCTs, "The World Bank is finally embracing science" (Deaton, 2010). Worrall (2007), Cartwright (2007) and Cartwright and Munro (2010) offer parallel critiques of RCTs in a general scientific context; Vandenbroucke (2004) argues for observational studies in medical research; Vandenbroucke et al. (2016) support "pragmatic pluralism" in establishing causality in epidemiology; and Ullman (2013, p. 37) calls on computer science to leave room for "ideas that require analysis rather than experiments".

[5] This led them to minimize its robust findings, which they summarize as "rejection of the null hypothesis that health spending does not respond to the out-of-pocket price"—a modest return on a very large investment.

[6] This is the first in a list of causal research questions, which Angrist and Pischke (2009) recommend to their readers as appropriate subjects for empirical research within the MEP. Like Projest STAR, empirical economic

decision-makers face, which is whether CSR is the optimal, or at least a good use of a large increase in education spending. Comparing the advantages of CSR to other options, or at least to its cost, requires a chain of evidence of which Project STAR can only provide a single link, where the other links in the chain are very weak, as we show in what follows. Tailoring the research question to maximize internal validity at the expense of policy relevance is a common feature of the MEP.

The narrowly defined *ceteris paribus*, experimental environment under which Project STAR was conducted ignores crucial differences between experimental conditions and the actual conditions of implementation. One important difference is the adverse general equilibrium effect that a statewide CSR is likely to have on the level and distribution of teacher quality; another is the reduced significance of improvements in test scores, when all test scores rise. In addition, Project STAR's randomized assignment of students and teachers to classrooms, while an effective control for selection bias, is a radical departure from most principals' purposeful, even strategic matching of teachers and students, which must affect the distribution of outcomes and may also affect their average level.

Moreover, the emphasis that the research question places on average class-size effects, veils the intrinsic heterogeneity of these effects. This is especially relevant for applying the lessons of Project STAR to prospective CSR initiatives in other settings. Ignoring how and when CSR affects learning reflects the theory-free anti-Bayesian bias of the counterfactual causal model underlying the MEP, which does not require an understanding of the mechanisms through which CSR benefits learning. However, as Cartwright and Munro (2010) argue, shifting the emphasis of policy-oriented research from determining whether a given social policy has a significant effect in specific circumstances to understanding the factors that determine its capacity to achieve such an effect can only enhance is usefulness as a guide to policy. Such an understanding could be furthered through economists collaborating with scholars in other disciplines with access to relevant prior knowledge, including "knowledge that we informally absorb" (Basu, 2013, p. 13). But this is rare; education scholars and economists have published extensively on Project STAR, but nearly always separately. Again, this is a general characteristic of research in the MEP.

---

research in this field invariably equates better learning with improved test scores, though test scores capture only a small part of learning, and even less of the other things schools are expected to do.

Finally, much of the economic research on Project STAR is uncommunicative. Policy is typically made in a contentious political and ideological context, where empirical findings and theoretical insights are routinely taken out of context or distorted by interested parties, if not happily ignored. The urge to take an unassailable evidence-based position is a natural response to these distortions. It generally leads to a heightened emphasis on the finer points of econometric analysis, rendered in technical language accessible only to a specialized audience of fellow practitioners that discourages dialog with the outside world. In the past, prominent economists engaged more readily in open debate, directly addressing the public on broad policy issues. This contributed to the adoption of many useful economic and social policies grounded in an eclectic evidence base.[7] The current disconnect is not only an obstacle to the wider communication of economic perspectives on policy but also undermines collaboration with practitioners, experts from other disciplines and decision-makers to design more useful empirical research.

The design and analysis of Project STAR

Project STAR (student/teacher achievement ratio) was commissioned by the State of Tennessee over thirty years ago, to determine whether classes should be reduced from their then current level of 22 students per class to a target level of 15 students per class (Folger, 1989; Finn and Achilles, 1990; Mosteller, 1995). In fall 1985, a non-random sample of 79 schools with 6,324 kindergarten children were chosen to participate in the project. Schools volunteered to participate in STAR for four years, from kindergarten to third grade, and had to be large enough to accommodate two large classes and one small class in the kindergarten cohort.[8] As the selection of schools was non-random, the research design focused on the estimation of class-size effects within schools.

In each of these schools, kindergarten children were randomly assigned to either a large class of 22-25 students, a large class with a full-time teacher's aide, or a small class of 13-17 students. In addition, small and large classes without a full-time teacher's aide had part-time teachers' aides. The initial randomization of students is described in detail by Folger (1989, p.

[7] Competition policy, the regulation of utilities, the shape of the tax code—among many examples—are based on ad hoc combinations of theoretical understanding and eclectic, incomplete empirical evidence. Worrall (2007) cites examples in medicine: the use of penicillin to cure pneumonia, and appendectomies.

[8] Of the 79 schools, four did not remain for the full 4 years, for unreported reasons (Hanushek, 1999, p. 151).

12) and Krueger (1999, note 7), and generally viewed as reliable, despite some apparent departures of implementation from design. The smaller classes have a slightly smaller share of disadvantaged children (Krueger, 1999, Table 1). Mathematics was mostly taught in the full class but reading was taught in groups, and remedial students were pulled out of class. Mosteller (1995, p. 124) notes that some 7% of children in small classes were found to be "incompatible" and moved to larger classes, but "[n]o mention is made of what was done about incompatible students who were already in regular-sized classes." Krueger, (1999, p. 499) notes that small classes actually ranged in size from 11 to 20 children, and regular classes from 15 to 30. Teacher assignment is described as random, but in a general way—there are no protocols—and there are indications that teachers' preferences might have been a factor in the assignment process.

Students were to study in these classes for four years, to the third grade, and tested annually in reading, mathematics and basic study skills, with the purpose of comparing average test scores across the different class types.[9] However, there was substantial attrition in the transition from kindergarten to the first grade, with 30% of the kindergarten cohort exiting the project. The first-grade sample was expanded to 6,828 students, 34% of them not having participated in the project the previous year. Moreover, as kindergarten was not mandatory in Tennessee, some 10% of the sample joining in first grade had not attended kindergarten at all. In addition, "… to alleviate some parental concerns, about half of the regular-class students in kindergarten were randomly reassigned to teacher-aide classes in first grade, and half of the teacher-aide pupils were reassigned to regular classes. Youngsters in small classes were not reassigned." (Finn and Achilles, 1990, p. 560).[10] Attrition was substantial in later years, too, and only about a third of the original kindergarten class continued through to the third grade. All told, 11,600 students participated in STAR over the four years of the program, with an average duration of 2.3 years. In all years, students exited the larger classes at a higher rate, and Ding and Lehrer (2010) present statistical evidence showing that attrition was non-random in all years.

The various estimates of class-size effects drawn from the Project STAR data generally agree that studying in a smaller class in kindergarten significantly improved test scores, with the

[9] Pate-Bain et al. (1992) and Nye et al. (1999) followed these students in later years, and subsequent studies followed them to college entry (Krueger and Whitmore, 2001) and the labor market (Chetty at al., 2011).

[10] As Krueger (1999, p. 499) observes "… if the constancy of one's classmates influences achievement, then the experimental comparison after kindergarten is compromised by the re-randomization".

estimated average effect falling in a fairly narrow range; and if there is an effect in later years it is much smaller. Hanushek (1999), a prominent critic of CSR, estimated a small-class effect in kindergarten of 0.17 of a standard deviation (SD), with possibly a small, added effect in first grade, and no further effect in grades two and three. Krueger (1999), a cautious supporter of CSR, grouped student-year effects by year-in-the-program. He found a first-year effect of 0.136 SD, and an average annual effect of 0.024 SD in each subsequent year, while noting that the second and third grade effects are smaller than the first-grade effect.[11] Ding and Lehrer (2010) applied a structural model to account for non-random attrition, and also found a significant effect in kindergarten and first grade, but none in later grades.

There is less agreement on the interpretation of these results, as there are various sources of potential bias in kindergarten, and even more so in subsequent grades. An upward bias might result if participating teachers and principals are eager for the experiment to succeed and take ad hoc steps to this effect (Hoxby, 2000, p. 1241);[12] or if teachers and students assigned in the experiment to larger classes are demoralized by their inferior conditions. Another possible source of bias is unobserved self-selection through differential attrition, which clearly occurred (Ding and Lehrer, 2010).[13] Alternatively, a downward bias could arise from measurement error, as a result of students moving in and out of classes during the year; or if teachers and students in larger classes make a greater effort to overcome their disadvantage; or if non-randomness in the allocation of teachers resulted in a larger proportion of experienced, more effective teachers assigned, say, to larger classes with a full-time teacher's aide. A detailed look at STAR highlights the obvious: the research design of a multi-year RCT in education, as in other areas of social policy, cannot be implemented with the precision of some RCTs in medical research.[14] There are multiple sources of possible bias in either direction, which can be assessed only roughly, and when the estimated effects are

[11] The first-year effect is the ratio of a class size effect of 2.99 + 0.65 = 3.64 percentiles (from Krueger, 1999, Table 9, column 3), to a standard deviation of 26.7 percentiles (from the Appendix Table).

[12] A possible example: the 7% of "incompatible" students in the smaller classes transferred to larger classes (Mosteller, 1995, p. 124; see above). This is distinct from either a "Hawthorn effect" or a "John Henry effect", which may also bias results; Krueger (1997, section E) partly controls for these.

[13] The direction of the bias is upward if more ambitious parents are pulling their children out of larger classes more frequently; or it might be downward, due to what Cameron and Heckman refer to as "dynamic selection bias", if CSR improves academic achievement across the board, and as a result a larger share of low-ability students in the control group do not continue with their cohort to the next grade level and fall out of the study.

[14] See the seminal contributions by Campbell (1957), Campbell and Stanley (1966) and Cook and Campbell (1979) on threats to validity in field experiments. "Double-blind" clinical medical trials avoid some of these problems, but are rarely an option in the social sciences. Of course, medical RCTs also experience problems of compliance and attrition, especially multi-year studies.

modest, as in the case of Project STAR, the precision of the econometric estimates may be illusory.[15]

While these limitations of Project STAR are significant, they do not represent the main weakness of RCTs as a guide to CSR policy. This stems, in the first instance, from Project STAR's research question. From a policy standpoint, the relevant question is not whether reducing class size in the early years of school, holding other factors constant, offers learning benefits; most parents and educators would answer that of course it does, at least within the range of class sizes examined by STAR (Folger, 1989, p. 123). The question, as Levin et al. (1984), Folger (1989), Harris (2002, 2007, 2009), Normore and Llon (2006), and others observe, is whether CSR is the best use of a large increase in education spending. One alternative use of extra funds is adding a teacher's-aid to larger classes, a line of inquiry originally included in STAR but compromised in the implementation; another is using the money to raise teacher salaries. But Project STAR was not designed to quantify the tradeoff between class size and teacher quality, as it responds to wage incentives, i.e., whether the benefits of a reduction in class size more than offset the implicit decline in teacher quality.

Krueger (1999, 2002a, 2003) considers the simpler question, whether the CSR examined by Project STAR was worth the expense, but even this cannot be answered adequately.[16] He assumes that costs are proportional to the number of classes, so that reducing class size from 22 to 15 students entails a cost increase of 7/15 = 47%, for 2.3 years, the average duration in the program. Krueger (2003) then estimates the economic value of learning gains from previous research on longitudinal data by Murnane, Willet and Levy (1995), Currie and Thomas (1999) and Neal and Johnson (1996), which he interprets as attributing an increase of 1.6% in lifetime income to the estimated gain of 0.20 standard deviation (SD) in third-grade scores. Allowing for variation in assumptions on discount rates and future wage growth yields an internal rate of return between 5.2% and 7.3%.[17]

---

[15] The cautious restatement by Aron-Dine et al. (2013) of the robust conclusions that may be drawn from the RAND health insurance experiment reflects a similar view; see also note 22, below.

[16] This corresponds to the analysis of the "flexible budget" case by Harris (2009), which suffers from the same flaws discussed here with reference to Krueger's analysis. Harris' analysis of the fixed budget case, which compares CSR to other interventions in terms of effect size per dollar, faces the complex challenge of finding a common denominator for comparing different effects.

[17] Even on its own terms, the link this calculation makes between gains in third grade tests and lifetime wages is heroic. As Card and Krueger (1996, footnote 2) observe: "… many studies find only a weak link between standardized tests and earnings" giving Murnane et al. (1995) as an example "[They] find that adding a math test score raises the R-squared by 1.7 percentage points for men and 4.0 percentage points for women."

However, this calculation ignores general equilibrium effects on both the costs and benefits of smaller classes that affect full-scale CSR initiatives.[18] Assuming a proportional 47% increase in costs to cover an increase of 47% in the demand for teachers leaves teachers' wages constant; and if the intra-marginal pre-CSR teachers are on average intrinsically better teachers than are the extra-marginal teachers (previously unemployed or employed in other professions) hired to meet new demand, then teacher quality declines. Krueger (2003, p. F59) does note in his list of caveats that "the quality of teachers could decline" as does Harris (2009), but neither factors this into their quantitative sensitivity analysis. Moreover, many of the newly hired teachers are likely to be inexperienced and uncredentialed, further lowering teacher quality in the short run. Jepsen and Rivkin's (2009) empirical analysis of the California CSR takes particular issue with the assumption of constant teacher quality, noting the significant effect of California's CSR on the quality and distribution of the teaching force, notably in the short term, with inexperienced teachers disproportionately employed in schools serving weaker populations.

General equilibrium effects also apply to the calculation of benefits. The estimated increase in wages attributed to higher test scores, in the longitudinal studies cited above, reflect the advantage of a higher test score to an *individual*. To the extent that this also represents a signaling effect, it is greater than the advantage to raising the test scores of an entire cohort, as positional gains are offset by positional losses; empirical evidence on the effect of test scores at the individual level is not a reliable guide to the value of a general increase.[19] Again, this is mentioned in the list of caveats but not incorporated quantitatively in the sensitivity analysis, and has no effect on Krueger's indicated range of rates of return. As Cartwright (2007) notes, the validity of evidence-based policy depends on the strength of the weakest link in the chain of evidence. Most analyses of Project STAR were less concerned with creating the strongest chain of evidence than with perfecting a single link in this chain.

Project STAR's focus on the average effect of class size on test scores raises a further set of issues. It glosses over the fundamental difference between a class-size effect that modestly

[18] Technically, these are departures from the "stable unit treatment value assumption" (SUTVA; Rubin, 2005, p. 323; Rothstein and von Wachter, 2016, p. 112). Imbens (2009, p. 2) notes that questions involving general equilibrium effects cannot be answered by simple experiments.

[19] This is similar to equating the average advantage of everyone scoring 100 points higher on their college entrance exams, to the advantage to an individual of gaining 100 points. Even a small state cannot ignore this consideration. Though its students might individually benefit from higher wages in a national labor market this would entail many of them moving out-of-state. And If all states implement CSR policies unilaterally to maximize net benefits, in a Nash equilibrium classes will be too small.

increases the scores of all students and one that greatly benefits a small number of stronger or weaker students.[20] Moreover, class-size effects estimated in other experimental and quasi-experimental studies vary widely, suggesting substantial heterogeneity in these effects. Thus, Hoxby's (2000) careful analysis of Connecticut data found no significant class-size effect, [21] and Cho et al. (2012) find very small effects in a study of class size effects in grades 3 and 5 in Minnesota. Shapson et al. (1980) find no class effects for most achievement measures analyzed in their randomized study of class size effects in Toronto schools, in the fourth and fifth grades. Angrist and Lavy (1999) exploit class-size caps in Israel, and find that a ten-student reduction in class size in the fourth and fifth grades increases achievement by 0.1-0.2 of a standard deviation but has no effect in the third grade.[22] Krueger (2002a, 2002b) and Hanushek (2002) drew different inferences from earlier non-randomized evidence.[23] This wide variety of estimates raises questions regarding the extent to which an average effect estimated from an intervention administered to a non-random set of Tennessee schools thirty years ago generalizes to other school populations. Understanding the sources of this heterogeneity could help answer this question, but it requires an understanding of the mechanisms through which CSR improves education outcomes.[24] This could be furthered by collaboration between economists and scholars from other disciplines with first-hand knowledge of teaching and learning processes but such collaboration is rare.[25]

[20] This is addressed, in a limited way, by looking at sub-populations, such as ethnic minorities, a departure from RCT protocol, which requires an adjustment in standard errors that is not always made. Moreover, taking an average of test scores, as a measure of learning, implies an underlying cardinal dimension where there is no such dimension. The estimated effect will depend on the specific design of the test, e.g., on the relative weighting of easier and more difficult subject material. This weighting is arbitrary, unless derived through some variant of Item Response Theory, the "Gold Standard" of test design, which determines question weights and student rankings simultaneously. But test scores so derived are inherently ordinal, and so can only rank students with respect to their performance on a common test or set of linked tests (Hambleton et al., 1991).

[21] Schanzenbach (2014, p. 5) sees the discrepancy between this result and the positive estimates from project STAR as an "unresolved puzzle". In drawing conclusions from RCTs, economic research attaches far less importance to repeated replication, as a condition for establishing validity, than do the natural sciences, a need emphasized in Campbell and Stanley's (1966) seminal analysis.

[22] See also Urquiola and Verhoogen (2009) for a critical look at their regression discontinuity design.

[23] This exchange led Krueger (2002b, p. 67) to restate his position more cautiously, maintaining that his is "not an attempt 'to provide a justification for undertaking large class size reductions'", but rather that based on the evidence, "one should be reluctant to conclude that school resources are irrelevant to student outcomes."

[24] Rothstein and von Wachter (2016) stress the importance of understanding the mechanisms behind the treatment effect. Ludwig et al. (2011) show how such an understanding can be used to design more useful and less costly experiments. Friedman-Sokuler (2016) argues for reframing empirical research to mirror the way policy is designed and implemented.

[25] What actually happens in the classroom is addressed in the education literature by Shapson et al. (1980) and Mitchell et al. (1989) among others. In the economics literature, McKee et al. (2015) examine class composition

A further weakness of STAR, as a means of informing public discussion of optimal class size is its use of randomization. To answer its specific, well-defined research question as accurately as possible, STAR randomizes the assignment of teachers and students to classrooms. But teachers and students are never randomly assigned to classrooms. School-district superintendents and school principals purposively assign teachers and students to schools and classrooms, in ways that vary from case to case, and of course teachers, students and parents also affect these assignments. Some principals may be maximizing average achievement where others strive to narrow achievement gaps; some superintendents may assign more-experienced teachers to more challenging schools where others may direct better-qualified teachers to schools serving more affluent and politically influential populations. Studying the impact of smaller classes under different assignment rules more closely approximates the actual conditions of implementation than random assignment.[26]

Other limitations of STAR as a guide to policy stem from differences between experimental conditions and full-scale implementation. One such difference is the opportunity that full-scale implementation affords to make complementary investments, such as investments in new teaching methods and materials. Reducing class size is the equivalent of adding inputs to a school's "production process"; knowing what to do with the extra inputs changes the production function. The original STAR design included minimal teacher training, the addition of a "modest" three-day training module between the second and third year in 15 selected schools, which had no significant effect on test results (Mosteller, 1995, p. 124). This lack of an effective training component in Project STAR can lead to underestimating the potential benefits of CSR.

These flaws had real effects. Project STAR was perceived as providing clear evidence of the economic value of CSR, and contributed to its subsequent enactment in many states (Hanushek, 2002). This is not to say that Project STAR was the main factor behind CSR adoption—a policy that was already widely popular among parents and educators before STAR (Folger, 1989, p. 123); but analyses of STAR that showed significant class-size effects helped persuade state legislators that the extra taxes they would have to raise for CSR would be money well spent. The largest and best documented of the state-level CSR initiatives in

---

in project STAR and find patterns consistent with reduced disruption playing a key role. They cite Lazear (2001) as an authority on class disruption, rather than primary sources in the education literature.

[26] In addition, Kasy (2016) shows that randomization may increase the expected mean-squared error. Bannerjee et al. (2016) discuss the excessive use of randomization in a broader context.

the years that followed was the California CSR, implementation of which began in 1996, at a recurring annual cost of $1.6 billion (Bohrnstedt and Stecher, 2002). It created extensive demand for new teachers, which more-affluent school districts were able to meet in a timely manner by attracting experienced, credentialed teachers from disadvantaged districts. Implementation of CSR lagged in districts serving disadvantaged minority and low-income students, and they were left with less-experienced teachers. Subsequent analysis by Bohrnstedt and Stecher (2002), Jepsen and Rivkin (2009) and Funkhouser (2009) found that much of the gains from reducing class size were offset by the decline in teacher quality, especially in disadvantaged districts. Consequently, the overall gains were smaller than those found in Project STAR, and the effect of CSR on the distribution of teacher quality widened the gap between advantaged and disadvantaged students. This runs counter to Project STAR's experimental findings, which indicated that smaller classes generate greater benefits for disadvantaged students.[27] Similar findings were reported for a large-scale CSR in Florida. Chingos' (2012, p. 543) concluded that "mandated CSR in Florida had little, if any, effect on student achievement"; and Normore and Llon (2006, p. 429) found that "reducing class sizes is the most expensive of state inputs that affect achievement scores."

Conclusion

This paper has focused on Project STAR, a particular field experiment in education, but the lessons it offers are more widely applicable to policy-oriented empirical studies in labor, health, welfare and related fields. It suggests several ways of enhancing the usefulness of RCTs and related research strategies, as a guide to policy.

- Optimize the tradeoff between methodological rigor and policy relevance.
- Align the research question with the policy issue; consider the strength of the full chain of evidence in deciding on which link to focus.
- Incorporate prior empirical knowledge in the analysis, including informal knowledge and knowledge from other disciplines.
- Shape the research to better understand the conditions under which intervention is most effective, and why, rather than whether it works in specific conditions.

[27] See, Krueger (2003,) and McKee et al. (2015) among others. Krueger (2003) does not refer to the California CSR; McKee et al. (2015) cite Jepsen and Rivkin (2009) on the California CSR without reference to its impact on distribution. Ludwig et al. (2011, p. 32) remark on this difference between experimental and actual outcomes.

- Incorporate the impact of general equilibrium effects on costs and benefits in actual implementation, in assessing the implications of estimated experimental effects.
- Recognize potential differences in the incentives faced by experimental subjects, compared to full-scale implementation, and differences in monitoring and control.
- Tailor the extent and form of randomization to conditions of actual implementation; full randomization may not be relevant for shaping policy.

Two points are worth further emphasis. The first is that the division of labor implicit in the MEP—academic economists perform experimental studies, leaving others to link their answers to wider policy issues—runs counter to what economists have traditionally been very good at: formulating and analyzing economic problems that involve optimizing the use of limited resources to achieve specified objectives. Economists are trained to define policy objectives and resource constraints, and relate them to measured quantities; recognize general equilibrium effects; account for incentives, and so on.

The second is that economic research, as McCloskey (1998) has been teaching us for decades, is a form of rhetoric. Effective rhetoric engages its audience. This means listening to the insights and concerns of elected and appointed officials, educators, parents, scholars from other disciplines and the general public, and formulating research in a way that relates to these concerns in terms that can be understood. Neglecting these essential analytical and rhetorical dimensions of research immediately diminishes its value as a basis for policy advice, and in the longer run may cause these skills to atrophy—to the loss of the profession and society at large.

References

Angrist, Joshua D. and Victor Lavy, 1999. Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement, *Quarterly Journal of Economics* 114:533-575.

Angrist, Joshua D. and Jörn-Steffen Pischke, 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Aron-Dine, Aviva, Liran Einav, and Amy Finkelstein, 2013. The RAND Health Insurance Experiment, Three Decades Later. *Journal of Economic Perspectives* 27(1):197–222.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg, 2016. Decision Theoretic Approaches to Experiment Design and External Validity. NBER Working Paper No. 22167.

Basu, Kaushik, 2013. The Method of Randomization, Economic Policy, and Reasoned Intuition. The World Bank Policy Research WP 6722.

Bohrnstedt, George W. and Brian M. Stecher, 2002. *What We Have Learned About Class Size Reduction in California*. CSR research consortium. Downloaded from: http://www.classize.org/techreport/CSRYear4_final.pdf

Cameron, Stephen V. and James J. Heckman, 1998. Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males. *Journal of Political Economy* 106(2): 262-333.

Campbell, Donald T., 1957. Factors relevant to the validity of experiments in social settings. *Psychology Bulletin* 54:297-312.

Campbell, Donald T., and Julian C. Stanley, 1966. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Card, David and Alan B. Krueger, 1996.Labor market effects of school quality: Theory and evidence. NBER Working Paper 5450.

Cartwright, Nancy, 2007. Are RCTs the Gold Standard? *Biosocieties* 2(1):11-20.

Cartwright, Nancy and Eileen Munro, 2010. The limitations of randomized controlled trials in predicting effectiveness. *Journal of evaluation in clinical practice* 16(2): 260-266.

Chetty, R., J. N. Friedman, N. Hilger, E.Saez, D. W. Schanzenbach and D. Yagan, 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics* 126, 4, pp. 1593—1660.

Chingos, Matthew, 2012. The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review* 31:543– 562.

Cho, Hyunkuk, Paul Glewwe and Melissa Whitler, 2012. Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review* 31:77– 95.

Cook, Thomas D., and Donald T. Campbell, 1979. *Quasi-Experimentation*. Chicago: Rand McNally.

Currie, J. and D. Thomas, 1999. Early test scores, socioeconomic status and future outcomes. NBER Working Paper 6943.

Deaton, Angus, 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* 48:424-455.

Deaton, Angus and Nancy Cartwright, 2016. Understanding and Misunderstanding Randomized Controlled Trials. NBER Working Paper 22595.

Ding, Weili and Steven F. Lehrer, 2010. Estimating Treatment Effects from Contaminated Multi-period Education Experiments: The Dynamic Impacts of Class Size Reductions. *The Review of Economics and Statistics* 92(1):31-42.

Finn, J. D. and Charles M. Achilles, 1990. Answers and questions about class size: a statewide experiment. *American Educational Research Journal*, vol. 27, pp. 557-77.

Folger, John, 1989. Lessons for Class Size Policy and Research. *Peabody Journal of Education* 67(1):123-132.

Friedman-Sokuler, Naomi, 2016. Empirical Economics, the Gold Standard and Public Policy: The Case of Class Size Reduction. Working paper. Ben Gurion University.

Funkhouser, Edward, 2009. The effect of kindergarten classroom size reduction on second grade student achievement: Evidence from California. *Economics of Education Review* 28: 403–414.

Ginsburg, Alan and Marshall S. Smith, 2016. Do Randomized Control Trials Meet the "Gold Standard"? A Study of the Usefulness of RCTs in the What Works Clearinghouse. Washington DC: American Enterprise Institute.

Greenberg, David, Mark Shroder, and Matthew Onstott, 1999. The social experiment market. *Journal of Economic Perspectives* 13(3):157–72.

Hambleton, R. K., H. Swaminathan, and H. J. Rogers, 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Hanushek, Eric. 1999. Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis* 21(2):143–64.

Hanushek, Eric. 2002. Evidence, politics, and the class size debate. In L. Mishel and R. Rothstein, eds., *The Class Size Debate*. Washington, DC: The Economic Policy Institute, pp. 37-65.

Harris, Douglas N., 2002. Identifying optimal class sizes and teacher salaries. In H. Levin and P. McEwan (Eds.), Cost Effectiveness Analysis in Education. Larchmont, NY: American Education Finance Association.

Harris, Douglas N. 2007. Class Size and School Size: Taking the Trade-Offs Seriously. *Brookings Papers on Education Policy*, No. 9, pp. 137-161.

Harris, Douglas N., 2009. Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis* 31(1):3-29.

Hausmann, Ricardo, 2016. "The Problem with Evidence-Based Policies." Project Syndicate,. http://prosyn.org/OOvVfVG

Heckman, James J. 2000. Causal Parameters and Pol icy Analysis in Economics: A Twentieth Century Retrospective. *Quarterly Journal of Economics* 115(1):45-97.

Heckman, James J. and Edward J. Vytlacil, 2007. Econometric evaluation of social programs, Part 1: causal models, structural models, and econometric policy evaluation. In James J. Heckman and Edward E. Leamer, eds., Handbook of Econometrics, 6B, 4779–874.

Hoxby, Caroline Minter, 2000. ''The Effects of Class Size on Student Achievement: New Evidence from Population Variation.'' *Quarterly Journal of Economics* 115(4):1239–86.

Imbens, Guido W., 2009. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). NBER Working Paper 14896

Jepsen, Christopher, and Steven Rivkin, 2009. Class size reduction and student achievement the potential tradeoff between teacher quality and class size. *Journal of Human Resources* 44: 223-250.

Kasy, Maximilian, 2016. Why experimenters might not want to randomize, and what they could do instead. *Political Analysis*, 1–15 doi: 10.1093/pan/mpw012

Krueger, Alan B., 1997. Experimental Estimates of Education Production Functions. NBER Working Paper 6051.

Krueger, Alan B., 1999. Experimental Estimates of Education Production Functions. *The Quarterly Journal of Economics* 114(2): 497-532.

Krueger, Alan B., 2002a. Understanding the magnitude and effect of class size on student achievement. In L. Mishel and R. Rothstein, eds., *The Class Size Debate*. Washington, DC: The Economic Policy Institute, pp. 7-35.

Krueger, Alan B., 2002b. Response to "Eric Hanushek's Evidence, politics, and the class size debate." In L. Mishel and R. Rothstein, eds., *The Class Size Debate*. Washington, DC: The Economic Policy Institute, pp. 67-87.

Krueger, Alan B., 2003. Economic considerations and class size. *Economic Journal* 113:F34–F63.

Krueger, Alan B. and Whitmore, Diane M., 2001. The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR. *The Economic Journal*, 111:468 (Jan), pp. 1-28.

Lancet, The, 2004. The World Bank is finally embracing science. Vol 364, p 731-2, Aug 28.

Lazear, Edward, 2001. Educational Production. *Quarterly Journal of Economics* 116(3): 777-803.

Levin, H., Glass, G., and Meister, G. (1984). Cost effectiveness of four educational interventions. (Project Report 84-All). Center for Education Research at Stanford.

Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan, 2011. Mechanism Experiments and Policy Evaluations. *Journal of Economic Perspectives* 25(3):17-38.

McCloskey, Deirdre N., 1998. *The rhetoric of economics*. University of Wisconsin Press.

McKee, Graham , Katharine R. E. Sims, Steven G. Rivkin, 2015. Disruption, learning, and the heterogeneous benefits of smaller classes. *Empirical Economics* 48:1267–1286.

Mitchell, Douglas E., Sara Ann Beach and Gary Badarak, 1989. Modeling the Relationship between Achievement and Class Size: A Re-Analysis of the Tennessee Project STAR Data. *Peabody Journal of Education* 67(1):34-74.

Mosteller, Frederick, 1995. The Tennessee Study of Class Size in the Early School Grades. *The Future of Children: Critical issues for children and youths* 5 (2) Summer/Fall.

Murnane, R., Willet, J. and Levy, F. 1995. The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* 77: 251–66.

Neal, D. and Johnson, W. 1996. The role of premarket factors in black-white wage differentials. Journal of Political Economy, vol. 104 (October), pp. 869–95.

Necker, Sarah, 2014. Scientific misbehavior in economics. *Research Policy* 43:1747–1759.

Normore, Anthony and Lynn Llon, 2006. Cost-Effective School Inputs: Is Class Size Reduction the Best Educational Expenditure for Florida? *Educational Policy* 20(2): 429-454.

Nye, Barbara, Larry V. Hedges, and Spiros Konstantopoulos, 1999 The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Education Evaluation and Policy Analysis* 21(2):127-142.

Pate-Bain, Helen , C. M. Achilles, Jayne Boyd-Zaharias and Bernard McKenna, 1992. Class Size Does Make a Difference. *The Phi Delta Kappan* 74(3):253-256.

Rothstein, Jesse and Till von Wachter, 2016. Social Experiments in the Labor Market. NBER Working Paper 22585.

Rubin, Donald B., 2005. Causal Inference Using Potential Outcomes, Journal of the American Statistical Association, 100:469, 322-331.

Schanzenbach, Diane Whitmore (2014). Does class size matter. Policy Briefs, National Education Policy Center, School of Education, University of Colorado, Boulder.

Shapson, Stan M., Edgar N. Wright, Gary Eason and John Fitzgerald, 1980. An Experimental Study of the Effects of Class Size. *American Educational Research Journal* 17(2):141-152

Sims, Christopher, 2010. But Economics Is Not an Experimental Science. *The Journal of Economic Perspectives*, 24(2), 59-68.

Ullman, Jeffrey D. "Experiments as research validation: Have we gone too far?" *Communications of the ACM* 58.9 (2015):37-39.

Urquiola, M and E. Verhoogen, 2009. Class-size caps, sorting, and the regression-discontinuity design. The American Economic Review, 99(1), 179-215.

Vandenbroucke, Jan P., Alex Broadbent, and Neil Pearce, 2016. Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology*. doi: 10.1093/ije/dyv341

Vandenbroucke, Jan P., 2004. When are observational studies as credible as randomised trials? *The Lancet* 363.9422: 1728-1731.

Worrall, John. 2007. "Evidence in Medicine and Evidence-Based Medicine" *Philosophy Compass* 2(6): 981-1022.