OXFORD REVIEW OF ECONOMIC POLICY, VOL. 20, NO. 2
DOI: 10.1093/oxrep/grh011

AMERICAN EDUCATION RESEARCH CHANGES TACK

JOSHUA D. ANGRIST
MIT and NBER[1]

For a quarter century, American education researchers have tended to favour qualitative and descriptive analyses over quantitative studies using random assignment or featuring credible quasi-experimental research designs. This has now changed. In 2002 and 2003, the US Department of Education funded a dozen randomized trials to evaluate the efficacy of pre-school programmes, up from one in 2000. In this essay, I explore the intellectual and legislative roots of this change, beginning with the story of how contemporary education research fell out of step with other social sciences. I then use a study in which low-achieving high-school students were randomly offered incentives to learn to show how recent developments in research methods answer ethical and practical objections to the use of random assignment for research on schools. Finally, I offer a few cautionary notes based on results from the recent effort to cut class size in California.

I. INTRODUCTION

The bread and butter issues in education research are questions about cause and effect in schools. Some of these questions concern the big picture of how schools are organized, such as whether school accountability schemes increase student achievement, or the many possible effects of school integration. Other questions are about types of teaching methods, such as whether computer-aided instruction increases test scores. Some questions touch on traditional economic concerns, such as whether the incorporation of financial performance incentives in teachers' work contracts benefits students. Finally, there are narrow questions about specific practices, such as whether a particular reading programme enhances reading comprehension. A unifying feature of these questions, and the feature that makes them—in my lexicon—'causal', is that they ask us to imagine alternative states of the world where groups of communities, schools, or students are differentially exposed to a particular intervention or reorganization. The causal effects of interest in these settings are differences in the outcomes we

[1] My thanks to the editors and to Howard Bloom and Alan Krueger for helpful comments on an earlier draft. Some of the work discussed in this paper was funded under NIH grant 1-R01-HD43809-01A1. I bear sole responsibility for the views expressed here.

would observe, given alternative policies or levels of exposure. Educators, policy-makers, and academics propose new interventions and policies in the hope that changes will improve on the current state of affairs.

A challenge facing people interested in causal questions is that these questions can never be answered with certainty. Two sorts of uncertainty limit our ability to determine, for example, whether a computer-assisted reading programme such as that studied by Rouse and Krueger (2004) increases test scores. First, there is sampling uncertainty. This arises from the fact that we are generally interested in drawing inferences for an entire population but typically we only have data for a sample. The scientific discipline of statistics is concerned with how to quantify and control this type of uncertainty. The relevant theory here is well understood and for the most part uncontroversial and easy to use.

A more fundamental difficulty is our inability to extract data of any sort from counterfactual states of the world. While we may be able to observe some students who were exposed to a new reading programme and others who were not, the difference in achievement levels between the students in these two groups is not necessarily due to the reading programme itself. The outcomes of non-exposed students provide a possible proxy measure for what outcomes would have been if exposed students had (counterfactually) never been exposed. But without controlling the process that determines exposure, we usually have little basis for presuming that the exposed versus non-exposed contrast provides an accurate picture of counterfactual differences. It may be—indeed, it seems likely—that students or teachers motivated to experiment with a new reading strategy are dissimilar in some key way from those not so motivated. On the one hand, they may be smarter or more ambitious. On the other, they may be doing poorly and searching for answers. In either case, the exposed versus non-exposed comparison provides a misleading measure of the causal effect of exposure since the groups being compared differ in a number of ways besides exposure.

The method of random assignment, invented by R. A. Fisher (1925) to study agricultural interventions when soil conditions are hard to measure and hold fixed, breaks the fundamental log-jam in the quest for counterfactual outcomes. When the classroom intervention or school reorganization plan of interest (from here on, I will call this the 'treatment') is randomly assigned, we can be confident that, at least on average, differences between groups with different exposure levels are due to the exposure itself and not to other uncontrolled factors. In other words, contrasts between randomly assigned treatments allow us to estimate the distribution of outcomes in counterfactual states (though we still need to worry about sampling uncertainty and perhaps a few other details). The formal argument behind this idea is most easily presented with a modest bit of mathematical notation, so I postpone this for now. The intuition is simply that random assignment balances everything in the treatment and control group except for the treatment itself. When treatment is randomly assigned, we can be confident that students in treatment and control groups have similar motivation, ability, and family background.

The logic behind the use of randomized experiments to estimate causal effects, sometimes also called randomized trials, has been found compelling enough to establish random assignment as the gold standard for causal inference in medicine and epidemiology (see, for example, Ioannidis *et al*., 2001) and, increasingly, in my own fields of labour economics and policy evaluation (see, for example, Glazerman *et al*., 2003). But not, at least for the past quarter century, in education.

II. THE EDUCATION RESEARCH LEGACY

In recent review articles, sociologist Thomas Cook (2001*a*,*b*) cast a jaundiced eye on the practice of education research. Cook's views are especially noteworthy since his textbook on research designs (Cook and Campbell, 1979) is widely used in disciplines ranging from psychology to economics.[2] In his review, Cook makes three key points: (i) in the 1960s and 1970s, education research used methods similar to those used in other social sciences, but this changed in the 1980s and 1990s; (ii) the shift in education research was due to the rise of a research

[2] The most recent version is Shadish *et al*. (2002).

ethos that reflects what Cook describes as a 'management consulting' view of programme evaluation; this view focuses on the organizational complexity of schools and gives pride of place to qualitative outcomes and flexible research designs, while eschewing systematic, quantitative evaluation that seeks general lessons; and (iii) quantitative and experimental education research continued on low boil after 1980, but scholarly work in this mould was almost always conducted by researchers working in fields other than education and outside academic schools of education.[3]

Cook's first point is that, at the dawn of the American social science renaissance in the 1960s and early 1970s, when many fields were fuelled by optimism about the potential for government action, education research appears to have been at the forefront of the trend towards systematic, quantitative, and often experimental social-policy evaluation. Drawing on Boruch *et al*.'s (2002) more detailed analysis, he makes explicit comparisons between education and the fields of criminology, social policy, and psychology. To this list I would add the discipline of economics, where participation in the quantitative social-science revolution is best exemplified by a series of negative income tax (NIT) experiments in which low-income families were offered different income-guarantee levels and benefit-withdrawal rates in an attempt to determine whether and by how much means-tested benefits were likely to affect work effort.

The NIT experiments have been extraordinarily influential in both scientific and policy terms, probably more so than social experiments in other fields. They created the foundation for an evaluation tradition that continues today in the form of multiple systematic evaluations of many welfare reform initiatives in American states.[4] The NIT experiments also stimulated the growth of a private-sector policy-evaluation industry, made up of a group of independent contractors such as Abt Associates, Mathematica Policy Research, The Manpower Demonstration Research Corporation, and Westat, with expertise in experimental design and the management of randomized trials. These organizations parallel in sophistication, if not in cash flow, the consultants who help pharmaceutical companies prepare new drug applications for the Food and Drug Administration.

Although this does not feature in Cook's discussion, a strong argument for the notion that modern education research was formed in the same quantitative crucible as criminology, economics, and psychology is the Perry pre-school project. The Perry project was a 1962 randomized trial of an early-intervention programme involving 123 black pre-schoolers in Ypsilanti (Michigan), about half of whom were randomly assigned to an intensive intervention that included pre-school education and home visits. It is hard to exaggerate the impact of this small study. Follow-up data were collected through 1993 when the participants were aged 27. Dozens of academic studies cite or use the Perry findings (see, for example, Barnett, 1992). Most importantly, the Perry project provided the intellectual basis for the massive Head Start intervention, begun in 1964, which has by now served millions of American children.

Why was the small-scale Perry study so influential? One possible answer is that it showed strong positive effects, always a welcome finding in the generally discouraging field of social-policy interventions. But a more likely explanation, since many programmes, at least superficially, seem beneficial, is that the Perry study appears to have been a well-designed randomized trial (see, for example, Schweinhart, 2003). Finally, it was, or at least has been until recently, one of few studies to use a randomized evaluation design for preschool education research. Currie's (2001) review of research on early childhood education programmes identifies only seven randomized trials and only three besides Perry that were well-designed (in the sense of suffer-

[3] See Burtless (2002) for a similar though somewhat more institutional view of the change in education research. Cronbach *et al.* (1980) provided the manifesto for Education Research as commonly practised. The introduction lays out 95 theses regarding programme evaluation, of which the 95th (p. 11) is: 'Scientific quality is not the principal standard; an evaluation should aim to be comprehensible, correct and complete, and credible to partisans on all sides.' Thesis 56 (p. 7) reveals the authors' far-reaching scepticism as to the possibility of scientific evaluation: 'Results of a program evaluation are so dependent on the setting that replication is only a figure of speech; the evaluator is essentially a historian.' This viewpoint leads to the question: why do any research at all?

[4] See Moffitt (2003) for a look back at these experiments and a discussion of how they led to contemporary 'in-work benefit' programmes such as the Earned Income Tax Credit in the USA and the Working Families Tax credit in the UK.

ing relatively little attrition and by virtue of following children at least into middle school). A forceful illustration of Cook's second point is that the much larger, long-running, and better-known Head Start programme has only now become the subject of a randomized evaluation (in an ongoing study).[5]

This is not to say that the last two decades' evaluation picture is entirely bleak. A bright light is the well-designed Tennessee class-size study (Finn and Achilles, 1990), probably best known through careful re-analyses by statisticians, Mosteller *et al*. (1996), and the economist, Krueger (1999*a*). More recently, school vouchers have been studied using random assignment (e.g. Howell and Peterson, 2002). Still, these studies are clearly the exception and, as Cook (2001*b*) notes, scholarly analyses using the data from these trials have come primarily from political scientists, psychologists, and economists.

A natural question at this point is why people such as Cook and me, aliens from another discipline, have been complaining about research methods outside our chosen fields. If researchers in education departments or schools of education prefer qualitative, flexible, less-than-scientifically rigorous evaluation, so be it. My response to this is a variant on the old saw about who gets to make strategic decisions for public policy: education research is too important to be left entirely to professional education researchers. In particular, the quality of education research is important for the same reason that the quality of medical research is important. In medicine, it is clear that misleading research results can be extraordinarily costly, both in dollars and in quality of life, sometimes even in terms of lives lost. The unfolding story of hormone replacement therapy (HRT) makes this case.

For decades women and their doctors believed that regular doses of estrogen and progestin could reduce bothersome post-menopausal symptoms and perhaps reduce the risk of heart attack. This belief is supported by theoretical reasoning and a large observational (i.e. non-randomized) study of nurses, some of whom used HRT and some who did not. The Nurses Health Study showed that hormone users had a 50 per cent reduced risk of heart disease. But a recently completed randomized trial showed that women randomly assigned to HRT were 29 per cent *more likely* to have had a stroke. Both studies are large enough that sampling uncertainty is not a major issue, but the randomized trial provides a better prediction of the long-term consequences of HRT than the nurses study because the women randomly assigned to take HRT are otherwise similar to the randomized controls. The trial results led the National Institutes of Health to issue an advisory against use of HRT in April 2004. The end of widespread HRT use is likely to save lives. Moreover, the resources devoted by patients and clinicians to this ineffective and perhaps even dangerous therapy can now be diverted to therapies likely to improve women's health without major side effects.

Of course, ineffective pedagogy or curricula rarely threaten students' lives in the direct way that harmful medical therapies can. But ineffective schools can certainly reduce the quality of students' lives since school quality is positively correlated with lifetime earnings, and may even shorten students' lives since education is associated with longevity.[6] Although I obviously picked a well-known medical example where results from a randomized study contradict the observational evidence, contradictions between the results of non-experimental studies and randomized trials appear in many fields.[7] It

[5] The contract for the National Head Start Impact Study was awarded in late 2002, with data to be collected from 2002 to 2006 (for more information, see http://www.acf.hhs.gov/programs/core/ongoing_research/hs/nhs_impact/nhs_impact.pdf). For a recent observational study of Head Start, see Garces *et al*. (2002). Although the intellectual and policy impact of the Perry study is undeniable, Gramlich (1986) sounds a sceptical note regarding some of the Perry findings, which show remarkably large improvements in outcomes for young adults in spite of the fact that the initial IQ gains for treated children disappeared within 2 years of treatment. Uncertainty as to the true effect of pre-school interventions should be reduced by the ongoing national Head Start study.

[6] For evidence on the quality–earnings relation, see Card and Krueger (1992). For recent evidence on the mortality link, see Lleras-Muney (2002). For more on HRT, see http://www.nhlbi.nih.gov/whi/

[7] In an influential study, Lalonde (1986) compared the results of an observational and randomized evaluation of a government training programme. See Glazerman *et al*. (2003) for an update. Bloom *et al.* (2002) compare experimental and non-experimental Welfare-to-Work evaluations. Ioannidis *et al.* (2001) present evidence that observational studies tend to overestimate the benefits of medical interventions. See Duflo and Kremer (2003) for a brief comparison of results from randomized and non-randomized studies in evaluations of economic development strategy and Weisburd *et al*. (2001) for a similar comparison in criminology.

therefore seems likely that education research is unexceptional in this regard, a point illustrated by the Rouse and Krueger (2004) study of educational technology mentioned in the introduction.

The Fast ForWord programme evaluated by Rouse and Krueger is a well-known and widely used computer-assisted reading intervention, motivated by a theory of how the brain processes sound. Previous evidence on Fast ForWord effectiveness comes largely from the researchers who developed and now market the programme, and consists primarily of before and after comparisons of programme users. Much of the evidence also consists of results for intermediate outcomes, such as brain function instead of actual reading ability. The Rouse and Krueger randomized trial shows no improvement in reading ability for Fast ForWord programme participants relative to a control group. The resources that school districts devote to this expensive and logistically complex programme would therefore likely be better allocated to other uses, where there is stronger evidence for increased learning.

III. EDUCATION RESEARCH GETS GAME

Of the 84 projects listed in the US Department of Education's (DOE) annual plan for Fiscal 2000, Boruch *et al*. (2002) identified only one as a randomized trial. In contrast, the web page listing awards for Fiscal 2002 and 2003 shows about a dozen awards supporting research designs using random assignment. These randomized studies focus primarily on pre-school and elementary-school curriculums. Even more remarkably, the document that solicited the grant proposals that led to this work *requires* proposed evaluation designs to use random assignment.[8]

The upsurge in randomized evaluations is not a fluke, but rather reflects a structural change in funding priorities at the DOE. What is driving this change? The impetus comes in part from an intellectual movement that cuts across the social sciences and sometimes unites those within disciplines who otherwise disagree. Among academic economists, for example, calls for renewed social experimentation have come from Alan Krueger (1999*b*), a well-known researcher who has written in favour of increased investment in schools, and from Eric Hanushek (2003), a widely cited school-spending sceptic.

Another important force for intellectual change has been the Campbell Collaboration, named after the late psychologist, Donald Campbell, of Cook and Campbell (1979) fame. The Campbell Collaboration brings together a diverse group of social scientists, including some working in schools of education, to advocate for and disseminate the results of randomized trials and high-quality research designs (for a description see *The Economist*, 2002). The Collaboration is managing the DOE's new What Works Clearinghouse, a web site designed to disseminate research findings quickly.[9] The very idea of such a clearinghouse, in which atomistic studies are catalogued for the sake of the general lessons they impart, is a powerful rejection of the 'organizational complexity' view of education research, which sees every school and every intervention as fundamentally unique.

Beyond intellectual foundations, a clear proximate cause of the randomization revolution is the 2001 *No Child Left Behind Act* (NCLBA) and especially the 2002 *Education Sciences Reform Act* (ESRA). The bombastic monicker notwithstanding, the NCLBA marks a significant and constructive break in the way the federal government views and supports education research. The NCLBA repeatedly calls for education policy to rely on a foundation of *scientifically based research*, defined as research using rigorous methodological designs and techniques, including control groups and, wherever possible, random assignment.[10]

The 2002 ESRA is a further expression of Congressional intent to reform American education research. The ESRA goes beyond the NCLBA by

[8] For details, see http://www.ed.gov/programs/edresearch/awards.html. The document referred to in this paragraph is known as a Request for Applications. This one is CFDA No. 84.305J.

[9] The clearinghouse can be found at http://www.w-w-c.org/index.html

[10] NCLB includes other important provisions that are less obviously constructive, such as sanctions for schools with low test scores. While routine standardized testing may be helpful for scientifically based research, in practice sensible accountability schemes are difficult to engineer. See Kane and Staiger (2002) for an analysis of the new accountability scheme.

changing the institutional framework through which education research is funded and creating the Institute of Education Sciences (IES). In a 2003 presentation to the American Educational Research Association (AERA), newly appointed IES Director Grover Whitehurst laid out the main features of the IES mission. These include an emphasis on results of practical value to educators and policy-makers and an emphasis on credible research designs, with randomized trials at the top of the methodological hierarchy.[11]

In the spirit of Cook (2001*a*) and Boruch *et al*. (2002), Whitehurst supported his critique of current research practice with a comparison of the contents of the AERA's leading scholarly journals to those of the *Journal of Educational Psychology* (JEP), published by the American Psychological Association. Over a 10-year period, only 6 per cent of studies in the AERA journals used random assignment, as compared with 48 per cent in the JEP. Similarly, 37 per cent of AERA articles used qualitative methods, while only 3 per cent of JEP articles used these methods. Finally, most of the non-research articles in AERA journals were pieces advocating a particular point of view, while most of the non-research pieces in JEP were traditional literature reviews or research syntheses. Whitehurst clearly intends, by force of the federal checkbook, to make the AERA research profile much closer to that in educational psychology. The official position is laid out in Institute of Education Sciences (2003).

IV. RANDOM ASSIGNMENT IN EDUCATION RESEARCH: PROBLEMS SOLVED AND REMAINING

In response to changing funding priorities at the DOE, the American Evaluation Association, a professional association that has been highly critical of IES initiatives, issued a press release. The release notes, among other things, that

> One of the reasons educational evaluation has moved beyond [the use of randomized trials], while not dismissing them entirely, is that very few educators consider it equitable to refuse participation to children who might be helped by a program, or to require their participation in an unproven program that takes them away from needed instruction.[12]

This statement, like others critical of the use of randomization in social research, raises two possible concerns. First, random assignment appears to require that some children or schools be denied services that may be beneficial. Second, random assignment appears to require that some children or schools be forced to participate in interventions of dubious merit. If true, these considerations would indeed generate practical and ethical problems. Luckily, however, random assignment can be used to produce powerful evidence on causal effects without the need either to deny services or compel participation. This possibility is an outgrowth of new developments in the methodology of causal research.

(i) A Pocket Primer on Research Methods

To explain how random-assignment research designs can be implemented without the need to deny services or compel participation, I use a little bit of mathematical notation and a specific example from my own work involving a randomized trial in Israeli high schools (Angrist and Lavy, 2002). The broad motivation for my study with Lavy is the fact that educators and policy-makers have long tried to reduce high-school drop-out rates. Most anti-drop-out interventions to date provide a range of remedial and support services for at-risk students. Unfortunately, randomized evaluations of service-oriented anti-drop-out strategies generally show these interventions to be ineffective (Dynarski and Gleason, 1998).

In Israel, few students drop out of high school. The most important transition for Israeli high-school students, and a hurdle that about half of all students fail to clear, is the attainment of a high-school matriculation certificate. Without a matriculation certificate, students cannot go to university. Thus, the role played by the matriculation certificate in Israel is very much like the role played by A levels in the UK. As in the USA, where attention focuses on reducing drop-out rates, Israel has looked to

[11] http://www.ed.gov/rschstat/research/pubs/ies.html
[12] Press release from 3 December 2003; see http://www.eval.org/doe.pressrelease.htm

service-oriented strategies to increase matriculation rates. The discouraging results from previous efforts stimulated our interest in a simpler approach that offers immediate financial rewards for student effort. As a theoretical matter, cash incentives may be helpful if low-achieving students discount the future very highly, reduce investment in schooling by going to work, or face peer pressure not to study. We therefore experimented with a pilot programme called Achievement Awards that provided substantial cash payments to low-achieving students who succeeded in passing their exams.[13]

The task of evaluating financial incentives in schools raises a number of practical and research-design questions, most importantly whether incentives should be offered to entire schools or to specific students within schools. Because results from a small pilot intervention suggested that random assignment to students within schools is unlikely to have an impact, we turned to a more ambitious intervention in which treatment was randomly assigned to entire schools. The school-based randomized trial was implemented with the cooperation of principals and administrators in treated schools.[14] In particular, our study sample consisted of 40 of Israel's worst high schools as measured by past matriculation rates, with half allocated randomly to the treatment group. The average matriculation rate in the sample of 40 was about 25 per cent in previous years. Beginning in June 2001, any student enrolled in a treated school who passed his or her matriculation exam was eligible for an award worth about $1,400.

A key feature of our research design, and an aspect that defuses the criticisms noted above, is that participation in our programme was voluntary in the treated group. In particular, to kick off the school-based demonstration, an orientation with principals and administrators from the 20 treatment schools was held in January 2001. We explained operational features of the programme and outlined the school-level requirements for award eligibility (school administrators had to provide us with a roster of enrolled students and to publicize the programme in their schools). Not surprisingly, some principals (three out of 20) chose not to participate. The principals of two other schools failed to provide the required enrolment roster in time. Thus, five out of 20 treated schools were non-compliant in the sense that they did not follow our intended treatment protocol. Of course, students in treated schools were also free to ignore the programme. For reasons explained below, non-compliance by principals or students does not compromise our experimental study design.

As noted in the introduction, the object of causal inference is a difference in counterfactual outcomes. To make this idea precise, let Y_{1j} denote the matriculation rate that would be observed in school j if the school were to be treated. Similarly, let Y_{0j} denote the matriculation rate that would be observed in school j if the school is not treated. Both outcomes are assumed to be meaningful for all schools even though only one is ever observed for a given school. In other words, we cannot hope to observed $Y_{1j}-Y_{0j}$, for a particular school, j. We might, however, hope to estimate the average difference in these two potential outcomes. This average difference as the average causal effect,

$$E[Y_{1j} - Y_{0j}] = E[Y_{1j}] - E[Y_{0j}],$$

where the symbol 'E', called the mathematical expectation operator, is shorthand for the 'average over all j', i.e.

$$E[Y_{1j}] = \frac{1}{40}\sum_{j=1}^{40} Y_{1j} \text{ and } E[Y_{0j}] = \frac{1}{40}\sum_{j=1}^{40} Y_{0j}.$$

To complete the notational framework, let the symbol D_j represent an indicator (dummy) variable that equals 1 when a school is treated and 0 otherwise. Note that we can now write the observed matriculation rate for school j as

$$Y_j = Y_{0j}(1-D_j) + Y_{1j}D_j. \quad (1)$$

This expression highlights the fact that only one of Y_{0j} and Y_{1j} is ever observed, since $D_j = 1$ implies 1–

[13] Our Achievement Awards programme has features in common with Britain's Education Maintenance Allowance (Ashworth *et al.*, 2001), though the latter has not been evaluated in a randomized trial. Also related is the experiment by Kremer *et al*. (2003), which used random assignment to evaluate achievement awards for primary school students in Kenya.

[14] Random assignment of schools rather than students generates a group-randomized trial (GRT) of the type widely used to study interventions in naturally clustered units such as schools, hospitals, and communities. GRTs offer practical advantages, but usually have lower statistical power than unclustered trials with the same sample size (see, for example, Feng *et al*., 2001).

$D_j = 0$ and vice versa. We can use the treatment dummy D_j to define another possible average causal effect, the effect of treatment in treated schools,

$$E[Y_{1j} - Y_{0j} | D_j=1] = E[Y_{1j} | D_j=1] - E[Y_{0j} | D_j=1],$$

where

$$E[Y_{1j} \big| D_j = 1] = \left\{ \sum_{j=1}^{40} Y_{1j} D_j \right\} / \left\{ \sum_{j=1}^{40} D_j \right\}$$

and

$$E[Y_{0j} \big| D_j = 1] = \left\{ \sum_{j=1}^{40} Y_{0j} D_j \right\} / \left\{ \sum_{j=1}^{40} D_j \right\}$$

are averages computed in the group of treated schools only. We can similarly define $E[Y_{1j} | D_j=0]$ and $E[Y_{0j} | D_j=0]$ as averages conditional on not being treated. Note that $E[Y_{1j} | D_j=1]$ and $E[Y_{0j} | D_j=0]$ are observed quantities, while $E[Y_{0j} | D_j=1]$ and $E[Y_{1j} | D_j=0]$ are counterfactual and cannot be directly observed.

Suppose initially that there is no random assignment, but rather that programme participation is fully voluntary. In other words, all 40 principals are given the opportunity to participate. It seems likely that principals who are better organized or have more motivated students would be more likely to take advantage of the opportunity to have their students win awards. If organizational ability or motivation translates into higher Y_{0j}, naïve comparisons of matriculation rates between participants and non-participants will produce a spurious positive treatment effect, even if the causal effect of treatment is really zero (in the sense that $Y_{1j}=Y_{0j}$ for every school). To see this, decompose the naïve comparison between those who do and do not participate as follows:

$$\begin{gathered} E[Y_j | D_j=1] - E[Y_j | D_j=0] = \\ E[Y_{1j} | D_j=1] - E[Y_{0j} | D_j=0] = \\ E[Y_{1j} | D_j=1] - E[Y_{0j} | D_j=1] + E[Y_{0j} | D_j=1] \\ - E[Y_{0j} | D_j=0] = \\ E[Y_{1j} - Y_{0j} | D_j=1] + \{E[Y_{0j} | D_j=1] - E[Y_{0j} | D_j=0]\}. \end{gathered}$$

The first term above is zero since I have postulated $Y_{1j}=Y_{0j}$ for every school. But the term in braces is likely to be positive since the principals who choose to participate have higher-than-average Y_{0j} while, conversely, those who do not choose to participate have lower-than-average Y_{0j}.

In this hypothetical scenario, the fact that $E[Y_{0j} | D_j=1] - E[Y_{0j} | D_j=0] \neq 0$ arises from the non-random process determining participation in the Achievement Awards programme. The process is such that participating schools differ from non-participating schools even in the absence of the programme. Now suppose, instead, that the decision to participate is determined by random assignment. Then the schools who do and do not get assigned to treatment should have roughly the same matriculation rates in the absence of treatment. In other words, with D_j randomly assigned, we have

$$E[Y_{0j} | D_j=1] - E[Y_{0j} | D_j=0] \approx 0.$$

In this case, we can expect the treatment–control difference to provide an unbiased estimate of the average treatment effect in treated schools.[15]

Neither of these scenarios (fully voluntary assignment, strict random assignment) corresponds to the research design used in Angrist and Lavy (2002). Rather, the Achievement Awards research design randomly assigned the *opportunity* to participate in the programme, with participation voluntary among those in the offered group. In medical trials, this sort of research design is said to randomly assign the *intention to treat*. To make the distinction between intention to treat and actual treatment status precise, let Z_j denote the former. That is, $Z_j=1$ if a principal is offered the opportunity to participate in Achievement Awards, while $D_j=1$ if he or she actually does participate.

Write $P[D_j=1 | Z_j=1]$ for the proportion of those with $Z_j=1$ who participate in the programme. This proportion is 15/20=0.75 since five out of 20 principals did not participate. On the other hand, $P[D_j=1 | Z_j=0]=0$, since no one who was not offered treatment participated. We can use these two facts to convert the intention-to-treat effect, i.e. the difference in average matriculation rates with Z_j switched

[15] If we also have $E[Y_{1j} | D_j=1] - E[Y_{1j} | D_j=0] \approx 0$, then this is the overall average treatment effect as well.

off and on, into the average causal effect on the treated. To see how, note that by a slight rearrangement of equation (1), we can write

$$Y_j = Y_{0j} + (Y_{1j}-Y_{0j})D_j. \qquad (1a)$$

Averaging matriculation rates in the offered group, we have

$$E[Y_j|\ Z_j=1] = E[Y_{0j}|\ Z_j=1] + E[(Y_{1j}-Y_{0j})D_j|\ Z_j=1].$$

In the not-offered group, we have the simpler expression

$$E[Y_j|\ Z_j=0] = E[Y_{0j}|\ Z_j=0]$$

since for everyone with $Z_j=0$ we know that $D_j=0$, i.e. those not offered do not participate.

The next step in the analysis of this research design is to notice that, by virtue of the random assignment of Z_j, those who were and were not offered the opportunity to participate should have similar Y_{0j}s, at least on average. Hence,

$$E[Y_{0j}|\ Z_j=1] - E[Y_{0j}|\ Z_j=0] \approx 0,$$

and therefore

$$E[Y_j|\ Z_j=1] - E[Y_j|\ Z_j=0] \approx E[(Y_{1j}-Y_{0j})D_j|\ Z_j=1].$$

Now, note also that

$$\begin{aligned} E[(Y_{1j}-Y_{0j})D_j|\ Z_j=1] &= 0 && \text{if } D_j=0 \\ &= E[Y_{1j}-Y_{0j}|\ D_j=1,\ Z_j=1] && \text{if } D_j=1, \end{aligned}$$

with the latter occurring in proportion $P[D_j=1|Z_j=1]$ of schools. Therefore,

$$E[(Y_{1j}-Y_{0j})D_j|\ Z_j=1] = E[Y_{1j}-Y_{0j}|\ D_j=1,\ Z_j=1]P[D_j=1|\ Z_j=1].$$

The argument is completed by observing that once I tell you that $D_j=1$, you also know that $Z_j=1$ since a school must have been offered treatment to take it. Since conditioning on $Z_j=1$ is redundant, $E[Y_{1j}-Y_{0j}|\ D_j=1,\ Z_j=1]=E[Y_{1j}-Y_{0j}|\ D_j=1]$. We have therefore shown

$$\{E[Y_j|\ Z_j=1] - E[Y_j|\ Z_j=0]\}/P[D_j=1|\ Z_j=1] \approx E[Y_{1j}-Y_{0j}|\ D_j=1]. \qquad (2)$$

Equation (2) captures one of the most important relationships in the theory of causal inference. It was originally derived by Howard Bloom (1984), an experienced programme evaluator who is now chief social scientist at the Manpower Demonstration Research Corporation. This equation says that in research designs where (*a*) the opportunity to participate is randomly determined, (*b*) no one in the control group receives treatment, and (*c*) treatment status is voluntary in the offered group, then the causal effect of treatment on the treated can be estimated by dividing the intention-to-treat effect by the compliance rate in the treated group. In the Angrist and Lavy study, for example, the difference in matriculation rates between the 20 schools offered the opportunity to receive Achievement Awards and the 20 schools not offered the opportunity to receive awards is 7.5 per cent. This is the intention-to-treat effect. Adjusting for the fact that only three-quarters of those offered treatment participated in the programme, the estimated effect of programme participation on participating schools is 7.5/0.75 or about 10 per cent.[16] This approach to causal inference addresses concerns about the need to compel experimental subjects to participate in exotic and/or unproven programmes.

What about the apparent need to deny services to members of the control group? In the Achievement Awards demonstration, no one in the originally selected 20-school control group had the opportunity to receive an award. But as far as the integrity of our research design goes, they could have if they wanted to. It turns out that non-compliance on the control group side is no more troubling than non-compliance in the treatment group. In a series of papers published over the last decade (e.g. Imbens and Angrist, 1994; Angrist *et al*., 1996), my co-authors and I have shown that causal effects can be estimated even when some of those in the control

[16] This is a sizable effect when measured against the control group mean of about 22 per cent, though it should be noted that this particular estimate is not significantly different from zero. Estimates for sub-samples of pupils with relatively high predicted matriculation rates are significant, however.

group receive treatment. Suppose, for example, that five schools in the Achievement Awards control sample had been able to worm their way into the treatment group, perhaps by lobbying Ministry of Education officials after random assignment. This would clearly corrupt the original random assignment in much the same way that non-compliance on the treatment side does.

When there are crossover subjects in both the treatment and control groups, dividing the intention-to-treat effect by the *difference* in compliance rates in the original treatment and control groups provides the appropriate adjustment. Thus, here we would divide by

$$P[D_j=1|\, Z_j=1]=0.75 \text{ minus } P[D_j=1|\, Z_j=0]=0.25,$$

which equals 0.5, instead of 0.75 when no one in the control group receives treatment. The resulting estimate, called a local average treatment effect (LATE), captures causal effects on the subset of the treated population that can be induced to take treatment simply by giving them an opportunity to do so. The LATE idea addresses concerns about the need to deny services. In fact, there is no such need, as long as the offer of treatment induces at least some subjects to receive a treatment that they would not have otherwise received.

The statistical procedure that accomplishes the appropriate adjustment for non-compliance is called an instrumental variables (IV) estimator, where the randomly assigned intention to treat is said to be 'an instrument' for endogenous or self-selected treatment status. The IV method, invented by the mathematician, economist (and poet!) Phillip Wright in the 1920s to estimate systems of simultaneous supply and demand equations, turns out to be an extraordinarily flexible tool that solves a number of significant methodological problems in causal research. Most importantly, IV methods correct for a range of compliance problems, including those arising in more complicated settings involving treatments of variable intensity such as hours of exam preparation or class size.[17]

(ii) Ethics and Random Assignment: Problem Solved?

The argument above shows how IV methods make it unnecessary either to deny services or compel participation in a randomized trial. At the same time, credible programme evaluation requires some scheme to induce differential rates of treatment in two otherwise comparable groups. Does this make randomized trials—even 'soft' randomized trials using instrumental variables—unethical? Not once we recognize two key facts about research and policy in the real world. First, and most importantly, in a world of finite budget constraints (all too finite in publicly funded education), access to services is *always* limited. In the American Head Start programme, for example, programme operators routinely deny services to some families because they simply cannot afford to include every interested family.

Second, if it is agreed that the purpose of causal research is an assessment of programme effects on outcomes, comparisons are inevitable. The only question is how these comparisons are to be drawn. As Cook and Payne (2002) note, without random assignment, the decision as to who gets treated is typically made on the basis of individual or bureaucratic judgements regarding need or merit. It seems unlikely that judgements about need or merit will be so clear cut that those making them can claim certainty. Once the uncertainty of these judgements is recognized, however, the essential fairness of allowing at least some uncertainty to enter in the process determining treatment assignment seems a logical conclusion. If we cannot make precise judgements about who most needs or merits an intervention, it seems only fair to allow those with equal claims an equal chance of access. Now, for 'equal chance of access', substitute 'random assignment'.

The ongoing evaluation of Britain's Education Maintenance Allowance (EMA) illustrates both of these considerations. The EMA aimed at improving educational outcomes by giving a weekly stipend to young adults from low-income families based on school attendance and achievement. Initially, funds

[17] For the use of IV to study class-size effects, see Angrist and Lavy (1999) and Krueger (1999*a*). For an introduction to IV, see Angrist and Krueger (2001). Stock and Trebbi (2003) give an intellectual history. Although the example in this section involves non-compliance on the part of principals (who can be seen as 'site administrators' in evaluation jargon), the logic of IV applies equally well to non-compliance among students.

were budgeted for a pilot study of EMA in only 15 of England's many Local Education Authorities (LEAs). In 2000, this was expanded to 56 LEAs, but this still included less than a third of young adults. As of September 2004, the EMA is to be available nationally, but for many years programme access was clearly limited.

The EMA pilot period provided an ideal opportunity to introduce an element of random assignment in the process determining LEA participation. The Department for Education and Skills (DfES), which administers the programme, might have put all urban, low-achieving LEAs into a randomly determined order, offering those at the top of the list the opportunity to participate first, with those lower down offered the chance to opt in later. Alternatively, the DfES might have offered the programme to *all* LEAs deemed eligible, but also have randomly offered some a financial incentive to wait a year or two. An experimental design using either of these schemes fits easily into an evaluation framework using IV methods. Instead, the DfES selected participant LEAs according to judgements as to whether they had high levels of deprivation, low participation rates in post-16 education, and low levels of attainment in year-11 examinations (Ashworth *et al.*, 2001). This is only one of many possible assignment schemes that would likely seem fair. And I bet there are a few dozen LEAs in Britain that fit this bill about equally well. Given that resources are limited and priorities uncertain, why not hold some sort of lottery among equally deserving areas?[18]

(iii) A Cautionary Tale from the Wild West

The modern theory of causal inference solves the most important practical and ethical problems in the use of random assignment for social experiments. This is not to say, however, that all problems are solved. The most important challenge to the use of random assignment is the question of external validity.

A research design that leads to an unbiased estimate of the causal effect of treatment in a particular study population is said to have internal validity. For example, the randomized trial studied in Angrist and Lavy (2002) is probably internally valid for the effect of Achievement Awards in Israel's worst schools. This is a useful result, since schools with low matriculation rates constitute the population of primary interest. That is, the purpose of this programme was to increase matriculation rates in just these sorts of schools. But the results of our study may not predict the effect of Achievement Awards should the programme be implemented more widely. A study that can be used to make predictions to another context is said to have external validity.[19] Other contexts include different populations and modified but related treatments.

The question of external validity is rarely clear-cut and cannot be resolved by improved statistical methodology. Rather, external validity comes from an understanding of the process linking treatments to outcomes. If we know why or how a treatment has a particular effect, then it seems reasonable to extrapolate past results to other settings where we believe the same forces are operating. On this point, I have to concede that there may be something to the notion of a useful sort of 'process evaluation' that is distinct from the 'impact evaluation' which I have been discussing so far.[20]

A recent example illustrates the role of external validity in determining the applicability of evaluation results. One of the most vexing and important questions in education research has been the effect of class size on student achievement. Our understanding of the *possible* benefits of smaller class size has been considerably enhanced by the Tennessee STAR randomized trial, which compared test scores between students randomly assigned to three groups: controls in regular classes with 22–25 students, a treatment group in regular classes with a teachers aid, and smaller classes with 13–17

[18] Mexico used just such a design to study an education incentive programme for rural communities known as Progresa (Schultz, 2004).

[19] For more on internal and external validity see Shadish *et al*. (2002).

[20] This is a minor concession on my part. An understanding of processes may help establish external validity but is not always necessary. Doctors have long prescribed effective interventions without the benefit of a full understanding of process. Moreover, without credible, i.e. internally valid, impact estimates that provide reliable information about programme consequences, it is hard to see why we should devote much effort to understanding programme process. For a recent study linking process and impact, see Bloom *et al*. (2003).

students. This unprecedented evaluation involved over 11,000 students in 80 schools and lasted 4 years. The results strongly suggest that smaller classes increase achievement, though a number of technical issues leave some room for interpretation as to the nature and magnitude of effects (see Krueger, 1999*a*).

Stimulated in part by the Tennessee results, in 1996, the state of California mandated a sharp drop in class size by financially rewarding schools that succeed in lowering average class size to 20 or fewer students in kindergarten and first, second, and third grades (roughly ages 5–8). At the time of the reform, average class size in California was around 30. The state initially committed to pay each school district $650 for every student in a grade where all students in the grade were in classes of 20 or fewer.[21] Although this is a substantial sum that led to a $1.3 billion payout in the programme's first year, it did not cover the average cost of class-size reduction in most districts. School districts quickly discovered that the cheapest way to reduce class size in two grades was often to combine them. Consider, for example, a district with 198 second-graders in 11 classes of 18 students each and 176 third-graders in 8 classes of 22 students each. Combining the two grades gives 19 classes with an average class size of 19.7, just under the required 20-student threshold, at no extra cost in classrooms or teachers.

The absence of a randomized trial makes it hard to assess the impact of the California programme. In a recently completed MIT Ph.D. thesis, however, Sims (2004) looked at the effect of California's class-size-reduction programme using a quasi-experimental research design similar to that used by Angrist and Lavy (1999). Sims finds that second- and third-grade students enrolled in districts where class size could be reduced most cheaply by combining classes were, indeed, likely to have had their grades combined. Overall, 12–15 per cent of second- and third-graders ended up in combination classes. The Sims results also suggest that students with an enrolment configuration that made cost savings from combination classes attractive suffered a sharp and statistically significant decline in test scores. The implication is that pooling students across grades is bad for learning. The adverse impact of combination classes is so large that, even if class-size reduction benefited the 85 per cent of students not in combination classes, the net effect of the California programme is probably negative, at least in the first years of the programme.

The Sims results use quasi-experimental variation based on the highly non-linear and even non-monotonic relation linking grade-specific enrolment patterns with the potential budget savings from combination classes. In particular, Sims uses the arcane details of this relation to develop a research design based on the 'regression-discontinuity method'. Although not as good as random assignment, the regression-discontinuity approach is usually much better than simply comparing participants and non-participants, since it implicitly compares treated subjects to an observational control group that is very similar (see, for example, Shadish *et al*., 2002). Taking Sims's results at face value, what is the moral of the California story? The Tennessee STAR randomized trial, which was undoubtedly part of the intellectual foundation for the California programme, led policy-makers to believe that smaller classes increase test scores. The end result, however, appears to be negative.

The moral, as I see it, is that the Tennessee results provide strong evidence that under relatively ideal circumstances (in particular, given long enough lead time and funding adequate to reduce class size without compromising other educational inputs), class-size reductions can increase test scores. This is an important result and a performance standard that many and perhaps most programmes, however well-intentioned, do not live up to. If the Tennessee experiment had shown no effects, the case for size-reduction proposals would clearly have been weakened. The Tennessee results do not establish, however, that class size reductions are always beneficial. In particular, they do not show that a modified intervention that combines the relatively proven strategy of smaller classes with the essentially untested (and, as it turns out, pernicious) strategy of pooling grades will be a net positive. This is simply too much external validity to ask for.

[21] The payment increased to $906 in 2002/3. Average per-pupil expenditure in California was $6,068 in 1995/6.

V. CONCLUSION: RANDOMIZATION IS NOT THE ONLY WAY BUT IT SETS THE STANDARD

As a researcher with a long-standing interest in the estimation of causal effects in economics, I have an instinctive sympathy for research designs using random assignment in any field of inquiry. At the same time, randomized trials, especially those requiring primary data collection, may be more expensive than quantitative studies that use existing surveys or administrative records without the benefit of random assignment. Clearly, there is some trade-off between the costs and benefits of alternative research designs. Non-randomized designs are most attractive when, without actually going to the trouble of doing the random assignment ourselves, we can find a source of natural variation that looks something like random assignment. The Sims (2004) and Angrist and Lavy (1999) studies discussed above are examples that find this variation in the essentially arbitrary nature of the bureaucratic process that determines class size. A similar idea is being used to evaluate Britain's Excellence in Cities intervention in inner city schools (for preliminary results, see Emmerson *et al*., 2004) and to evaluate a British reading programme called The Literacy Hour (Machin and McNally, 2003).

Another study design that mimics random assignment uses lotteries as a source of variation in treatment assignment. Typically, these lotteries were established for purposes of fairness and not as a research tool, but this fact makes them no less useful for research. Two recent evaluations of this sort look at the effect of school vouchers (Rouse, 1998; Angrist *et al.*, 2002). A distinguishing feature of both regression-discontinuity research designs and lottery-based quasi-experimental designs is that the force of evidence they generate can easily be judged according to whether these designs do in fact, 'provide a good experiment'. In other words, the notion of a properly executed randomized trial provides the appropriate benchmark for this judgement, in education as in other social sciences, even when a randomized trial is out of reach.

In closing, a brief cautionary note seems warranted. Although recent developments leave me optimistic, it is premature to declare victory in the struggle over education research methodology. The most significant threats to the randomization camp appear now to come from within. I am particularly concerned about the Trojan horses of conflicts of interest, poor data collection, and complex experimental designs that ultimately compromise the integrity of a trial.[22] Because randomized trials have the veneer of science, even when they are poorly executed, lapses in these areas will tend to discredit scientifically based research more generally. My hope is that these threats can be countered with the usual weapons of arm's-length evaluations, careful replication, and peer review. Assuming these threats are neutralized, the new emphasis on scientific research designs should allow education research to take its natural place as a public good benefiting all those with a stake in the quality of our schools.

REFERENCES

Angrist, J. D., and Krueger, A. (2001), 'Instrumental Variables and the Search for Identification', *Journal of Economic Perspectives*, **15**(Fall), 69–86.

— Lavy, V. (1999), 'Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement', *Quarterly Journal of Economics*, **114**(2), 533–75.

— —(2002), 'The Effect of High School Matriculation Awards: Evidence from Randomized Trials', NBER Working Paper 9389, December.

— Imbens, G. W., and Rubin, D. B. (1996), 'Identification of Causal Effects Using Instrumental Variables', *Journal of the American Statistical Association*, **91**(434), 444–55.

— Bettinger, E., Bloom, E., King, E., and Kremer, M. (2002), 'Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment', *American Economic Review*, **92**(December), 1535–58.

[22] This last problem bedevils analysts of the Seattle–Denver NIT experiments (Ashenfelter and Plant, 1990) and appears more recently in the New York City voucher experiment (Barnard *et al.*, 2003; Krueger and Zhu, 2003).

Ashenfelter, O. C., and Plant, M. (1990), 'Non-parametric Estimates of the Labor Supply Effects of Negative Income Tax Programs', *Journal of Labor Economics*, **8**(2), S396–415.

Ashworth, K., *et al.* (2001), *Education Maintenance Allowance: The First year, A Quantitative Evaluation*, Department for Education and Evaluation Research Brief 257, May.

Barnard, I., Frangakis, C. E., Hill, J. L., and Rubin, D. B. (2003), 'Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Vouchers in New York City', *Journal of the American Statistical Association*, **98**, 320–3.

Barnett, W. S. (1992), 'Benefits of Compensatory Education', *Journal of Human Resources*, **27**(Spring), 279–312.

Bloom, H. S. (1984), 'Accounting for No-shows in Experimental Evaluation Designs', *Evaluation Review*, **8**, 225–46.

— Michalapoulos, C., Hill, C., and Lei, Y. (2002), 'Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?', MDRC Working Paper on Research Methodology, New York, MDRC, June.

— Hill, C. J., and Riccio, J. A. (2003), 'Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments', *Journal of Policy Analysis and Management*, **22**(4), 551–75.

Boruch, R., De Moya, D., and Snyder, B. (2002), 'The Importance of Randomized Field Trials in Education and Related Areas', in F. Mosteller and R. Boruch (eds), *Evidence Matters: Randomized Trials in Education Research*, Washington, DC, The Brookings Institution.

Burtless, G. (2002), 'Randomized Field Trials for Policy Evaluation: Why Not in Education?', in F. Mosteller and R Boruch (eds), *Evidence Matters: Randomized Trials in Education Research*, Washington, DC, The Brookings Institution.

Card, D. E., and Krueger, A. B. (1992), 'Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States', *Journal of Political Economy*, **100**.

Cook, T. (2001*a*), 'Sciencephobia: Why Education Researchers Reject Randomized Experiments', *Education Next* (www.educationnext.org), Fall, 63–8.

— (2001*b*), 'A Critical Appraisal of the Case Against Using Experiments to Assess School (or Community) Effects', http://www.educationnext.org/unabridged/20013/cook.html

— Campbell, D. T. (1979), *Quasi-experimentation: Design and Analysis Issues for Field Experimentation*, Chicago, IL, Rand-McNally.

— Payne, M. (2002), 'Objecting to the Objections to Using Random Assignment in Educational Research', ch. 6 in F. Mosteller and R. Boruch (eds), *Evidence Matters: Randomized Trials in Education Research*, Washington, DC, The Brookings Institution.

Cronbach, L., *et al.* (1980), *Towards a Reform of Program Evaluation: Aims, Methods, and Institutional Arrangements*, San Francisco, CA, Jossey-Bass.

Currie, J. (2001), 'Early Childhood Education Programs', *Journal of Economic Perspectives*, **15**(Spring), 213–38.

Duflo, E., and Kremer, M. (2003), 'Use of Randomization in the Evaluation of Development Effectiveness', paper prepared for the World Bank Operations Evaluation Department, Washington, DC, July.

Dynarski, M., and Gleason, P. (1998), *How Can We Help? What Have We Learned from Evaluations of Federal Dropout-Prevention Program*, Princeton, NJ, Mathematica Policy Research Report 8014-140, June.

Emmerson, C., McNally, S., and Meghir, C. (2004), 'Economic Evaluation of Education Initiatives', ch. 9 in S. Machin and A. Vignoles (eds), *What's the Good of Education? The Economics of Education in the United Kingdom*, forthcoming, Princeton, NJ, Princeton University Press.

Feng, Z., Diehr, P., Peterson, A., and McLerran, D. (2001), 'Selected Statistical Issues in Group Randomized Trials', *Annual Review of Public Health*, **22**, 167–87.

Finn, J. D., and Achilles, C. M. (1990), 'Answers and Questions about Class Size: A Statewide Experiment', *American Educational Research Journal*, **27**(3), 557–77.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh, Oliver & Boyd.

Garces, E., Thomas, D., and Currie, J. (2002), 'Longer-term Effects of Head Start', *American Economic Review*, **92**(September), 999–1012.

Glazerman, S., Levy, D. M., and Myers, D. (2003), 'Nonexperimental versus Experimental Estimates of Earnings Impacts', *The Annals of the American Academy of Political and Social Science*, **589**(September), 63–93.

Gramlich, E. M. (1986), 'Evaluation of Education Projects: the Case of the Perry Preschool Program', *Economics of Education Review*, **5**(1), 17–24.

Hanushek, E. (2003), 'Comment', in B. Friedman (ed.), *Inequality in American: What Role for Human Capital Policies*, Cambridge, MA, MIT Press, 252–69.

Howell, W. G., and Peterson, P. E. (2002), *The Education Gap: Vouchers and Urban Schools*, Washington, DC, The Brookings Institution.

Imbens, G. W., and Angrist, J. D. (1994), 'Identification and Estimation of Local Average Treatment Effects', *Econometrica*, **62**(2),467–75.

Institute of Education Sciences (2003), 'Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User-friendly Guide', Washington DC, US. Department of Education, available at http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

Ioannidis, J. P., *et al.* (2001), 'Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies', *Journal of the American Medical Association*, **286**(7), 821–30.

Kane, T., and Staiger, D. (2002), 'The Promise and Pitfalls of Using Imprecise School Accountability Measures', *Journal of Economic Perspectives*, Fall.

Kremer, M., Miguel, E., and Thornton, R. (2003), 'Incentives to Learn', Harvard Department of Economics, mimeo, October.

Krueger, A. B. (1999*a*), 'Experimental Estimates of Education Production Functions', *Quarterly Journal of Economics* **114**, 497–532.

— (1999*b*), 'But Does it Work?', *The New York Times*, 7 November.

— Zhu, P. (2003), 'Comment', *Journal of the American Statistical Association*, **98**, 314–17.

Lalonde, R. J. (1986), 'Evaluating the Econometric Evaluations of Training Programs Using Experimental Data', *American Economic Review*, **76**(4),602–20.

Lleras-Muney, A. (2002), 'The Relationship Between Education and Adult Mortality in the United States', NBER Working Paper 9185, September.

Machin, S., and McNally, S. (2003), 'The Literacy Hour', London School of Economics, Center for the Economics of Education, mimeo, December.

Moffitt, R. A (2003), 'The Negative Income Tax and the Evolution of US Welfare Policy', NBER Working Paper 9751, June.

Mosteller, F., Light, R. J., and Sachs, J. A. (1996), 'Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size', *Harvard Educational Review*, **66**, 797–842.

Rouse, C. E. (1998), 'Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program', *Quarterly Journal of Economics*, **13**(2), 553–602.

— Krueger, A. B. (2004), 'Putting Computerized Instruction to the Test: A Randomized Evaluation of a "Scientifically-based" Reading Program', *Economics of Education Review*, forthcoming.

Schultz, T. P. (2004), 'School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program', *Journal of Development Economics*, **74**(June), 199–250.

Schweinhart, L. J. (2003), 'Benefits, Costs, and Explanation of the High/Scope Perry Preschool Program', paper presented at the Meeting of the Society for Research in Child Development, Tampa, Florida, 26 April.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Boston, MA, Houghton Mifflin.

Sims, D. (2004), 'How Flexible is Education Production? Combination Classes and Class Size Reduction in California', MIT Department of Economics, mimeo, January.

Stock, J., and Trebbi, F. (2003), 'Who Invented Instrumental Variables Regression?', *Journal of Economic Perspectives*, **17**, 177–94.

The Economist (2002), 'Try it and See', 28 February, Print Edition.

Weisburd, D., Lum, C., and Petrosino, A. (2001), 'Does Research Design Affect Study Outcomes in Criminal Justice?', *Annals of the American Academy of Social and Political Sciences*, **578**(November), 50–70.