Education Economics

ISSN: 0964-5292 (Print) 1469-5782 (Online) Journal homepage: www.tandfonline.com/journals/cede20

Econometric methods for causal evaluation of education policies and practices: a non-technical guide

Martin Schlotter, Guido Schwerdt & Ludger Woessmann

To cite this article: Martin Schlotter, Guido Schwerdt & Ludger Woessmann (2011) Econometric methods for causal evaluation of education policies and practices: a non-technical guide, Education Economics, 19:2, 109-137, DOI: 10.1080/09645292.2010.511821

To link to this article: https://doi.org/10.1080/09645292.2010.511821

Published online: 04 Jan 2011.

Education Economics
Vol. 19, No. 2, May 2011, 109–137

Econometric methods for causal evaluation of education policies and practices: a non-technical guide

Martin Schlotter, Guido Schwerdt and Ludger Woessmann*

Ifo Institute for Economic Research, University of Munich, Munich, Germany

(*Received 24 November 2009; final version received 29 March 2010*)

Education policy-makers and practitioners want to know which policies and practices can best achieve their goals. But research that can inform evidence-based policy often requires complex methods to distinguish causation from accidental association. Avoiding econometric jargon and technical detail, this paper explains the main idea and intuition of leading empirical strategies devised to identify causal impacts and illustrates their use with real-world examples. It covers six evaluation methods: controlled experiments, lotteries of oversubscribed programs, instrumental variables, regression discontinuities, differences-in-differences approach, and panel data techniques. Illustrating applications include evaluations of early childhood interventions, voucher lotteries, funding programs for disadvantaged students, and compulsory school and tracking reforms.

JEL classification: I20; C01

Keywords: causal effects; education; policy evaluation; non-technical guide

1. Introduction: how to learn what works

1.1. Evidence-based policy: motivation and overview

The need for evidence-based policy in the field of education is increasingly recognized (e.g., Commission of the European Communities 2007). However, providing empirical evidence suitable for guiding policy is not an easy task, because it refers to causal inferences that require special research methods which are not always easy to communicate due to their technical complexity. This paper surveys the methods that the economics profession has increasingly used over the past decade to estimate effects of educational policies and practices. These methods are designed to distinguish accidental association from causation. They provide empirical strategies to identify the causal impact of different reforms on any kind of educational outcomes.

The paper is addressed at policy-makers, practitioners, students, and researchers from other fields who are interested in learning about causal relationships at work in education but are not familiar with modern econometric techniques. Among researchers, the exposition is not aimed at econometricians who use these techniques, but rather at essentially any interested non-econometrician – be it theoretical or macro-economists or non-economist education researchers. The aim is to equip the interested reader with the intuition of how recent methods for causal evaluation

*Corresponding author. Email: woessmann@ifo.de

ISSN 0964-5292 print/ISSN 1469-5782 online

DOI: 10.1080/09645292.2010.511821
http://www.informaworld.com

work and to point out their strengths and caveats. This will not only facilitate the reading of recent empirical studies evaluating educational policies and practices but also enable the reader to interpret results and better judge the ability of a specific application to identify a causal effect. To do so, this paper provides a guide to the most recent methods that tries to circumvent any econometric jargon, technicality, and detail.[1] Instead, it discusses just the key idea and intuition of each of the methods and then illustrates how each can be used by a real-world example study based on a successful application of the method, with a particular focus on European examples.

It is, however, useful to note that the methods described here are by no means confined to the economics profession. In fact, it was the American Educational Research Association, with its broad range of interdisciplinary approaches to educational research in general, which recently published an extensive report on *Estimating Causal Effects using Experimental and Observational Designs* (Schneider et al. 2007) which in large parts deals with the same kind of methods described here. The issues and methods discussed in this paper basically apply to any quantitative evaluation of policies.

To prepare the ground, the remainder of this introductory section presents the central issue and points out the limitations of standard classical methods that build on the idea of controlling for observable factors. Against this background, the paper starts its presentation of methods for causal evaluation with two methods based on explicit randomization – controlled experiments (Section 2.1) and randomized lotteries of oversubscribed programs (Section 2.2). We then turn to two methods that researchers use to emulate experimental settings using observational data in what might be called natural experiments – the instrumental-variable (IV) approach (Section 3.1) and the regression-discontinuity approach (Section 3.2). It closes with two panel data-based methods that aim to account for endogeneity in observational data – the differences-in-differences approach (Section 4.1) and additional panel data techniques (Section 4.2). These six groups of evaluation methods are illustrated with a wide range of example applications, including evaluations of an early childhood intervention program, a voucher lottery, the effects of extended education on labor market outcomes, a program that provided extra funding for disadvantaged students, a reform that increased the compulsory schooling age and abolished early tracking, and the effects of peers on student outcomes.

1.2. The issue: from correlation to causation

Using standard statistical methods, it is reasonably straightforward to establish whether there is an association between two things – for example, between the introduction of a certain education reform (the treatment) and the learning outcome of students (the outcome). However, whether such a statistical correlation can be interpreted as the causal effect of the reform on outcomes is another matter. The problem is that there may well be other reasons why this association comes about.

For example, if schools with relatively poor student outcomes are induced to implement the reform, then we may well find a negative association between student outcomes and the reform across schools. But the reason for this association is a causal effect of poor student outcomes on implementing the reform, not the other way round. This is an example of ‘reverse causality’ where the outcome of interest actually asserts a causal effect on the treatment of interest.

Another example of alternative reasons for the association is that of 'omitted variables', where a third variable affects both treatment and outcome. Think of differences in the preferences for high-quality education among parents. If some parents value education more than others, then they will be more inclined to lobby for smaller classes and will also support their children in many other dimensions, for example, by buying them afternoon classes. If these other dimensions improve the learning outcomes of their children, we will observe a positive association between smaller classes and student outcomes. But we would observe this even if there were no causal effect whatsoever of class size on student outcomes: The third variable of parental preferences gives rise to the association. If this so-called 'unobserved heterogeneity' between the treated and the untreated individuals refers to variables that affect both the treatment and the outcome, identification of causal effects will be hampered.

Whenever other reasons exist that give rise to some correlation between the two things of interest – the treatment and the outcome – the overall correlation cannot be interpreted as the causal effect of the treatment on the outcome. Broadly speaking, this is what economists call the 'endogeneity problem'. The term stems from the idea that treatment cannot be viewed as exogenous to the model of interest, as it should be, but that it is rather endogenously determined within the model – depending on the outcome or being jointly determined with the outcome by a third factor. Because of the problem of endogeneity, estimates of the association between treatment and outcome based on correlations will be biased estimates of the causal effect of treatment on outcome.[2]

Standard approaches try to deal with this problem by observing the other sources of possible correlation and take out the difference in outcomes that can be attributed to these other observed differences. This is the approach of multivariate models that estimate the effects of multiple variables on the outcome at the same time, such as the classical ordinary least-squares (OLS) or multilevel modeling (or hierarchical linear models, HLM) techniques. They allow estimating the association between treatment and outcome conditional on the effects of the other observed factors. In the omitted-variable example, if we can fully observe the omitted variable, we just add it to the multivariate model and the confounding effect is properly captured. But more often than not, we cannot fully observe the omitted variable: In the example, how would we get a perfect measure of parents' valuation of high-quality education? But as long as part of the omitted variable stays unobserved, the estimated conditional association will not necessarily warrant a causal interpretation.

It becomes obvious that whenever relevant factors that are associated with both treatment and outcome remain unobserved, the classical methods have to be interpreted with great care. Over the past decades, it has become increasingly apparent in the literature evaluating education policies and practices that there are myriad important factors that remain unobserved in our models, often rendering the attempts to control for all relevant confounding factors in vain. Just think of such factors as the innate ability of students, parental preferences for certain outcomes, the teaching aptitude of teachers, or the norms and values of peers and neighborhoods. Even if we manage to obtain observable measures of certain dimensions of these factors, others – often important ones – will remain unobserved. Even more, controlling for observable factors does not solve the endogeneity problem when it is due to plain reverse causality in that the outcome causes the treatment. The only solution is to search for variation in treatment that is not related with other factors that are correlated with the outcome.

The same caveats that apply to the classical models also apply to another technique that has recently become increasingly popular, namely the matching technique. The central idea of this technique is to find matching pairs of treated and untreated individuals who are as similar as possible in terms of observed (pre-treatment) characteristics. Under certain assumptions, this method can reduce the bias of the treatment effect. But as long as relevant factors remain unobserved, it cannot eliminate the bias (see, e.g., Becker and Ichino 2002). In this sense, the matching technique cannot solve the endogeneity problem and suffers as much from bias due to unobserved factors as the classical models.[3]

Given the limitations of classical and matching methods, in this paper, we turn to new techniques, increasingly applied by economists in recent years, that aim to provide more convincing identification of causal effects in the face of unobservable confounding factors. In medical trials, only some patients get treated, and the assignment to the group of treated and untreated patients is done in a randomized way to ensure that it is not confounded with other factors. The untreated patients are then used as a so-called control group with which the treated patients are compared. The aim of the new techniques is to mimic this type of experimental design, often using data not generated by an explicitly experimental design. The techniques aim to form a treatment group (that is subject to the treatment) and a control group (that is not subject to the treatment) which are exactly the same. That is, they should not have been subdivided into treatment and control group based on reasons that are correlated with the outcome of interest. Ideally, we would like to observe the same individuals at the same point in time both in the treated status and in the untreated status. Of course, this is impossible, because the same individual cannot be in and out of treatment at once. Therefore, the key issue is estimating what would have happened in the counterfactual situation – which outcome a treated individual would have had if she had not been treated.

The central idea of the new techniques is that if the separation of the population into treatment and control group is purely random and a sufficiently large number of individuals are observed, then randomness ensures that the two groups do not differ systematically on other dimensions. In effect, the mathematical law of large numbers makes sure that the characteristics of those in the treatment group will be the same as those in the control group. Thus, the causal effect of the treatment on the outcome can be directly observed by comparing the average outcomes of the treatment group and the control group, because the two groups differ only in terms of the treatment. The aim of the evaluation methods discussed in this paper is to generate such proper treatment and control groups and thus rule out that estimates of the treatment effect are biased by unobserved differences.

2. Tossing the dice: explicit randomization

We start with two techniques that use explicit randomization to build proper treatment and control groups that do not differ from one another in observed or unobserved dimensions.

2.1. The ideal world: controlled experiments

The first technique is to implement a randomized experiment. Although some people may have reservations with experimenting when it comes to social settings, the only important feature about controlled experiments is the random assignment of participants to treatment and control groups. From a conceptual perspective, such controlled experiments constitute an important benchmark with certain ideal features against which to judge the other techniques discussed in this paper. Ethical issues of conducting experiments are discussed below.

2.1.1. Idea and intuition

In order to conclude that a specific educational reform or program has a causal effect on economic or social outcomes of individuals who participated in the intervention, we would, ideally, like to observe the same persons in a counterfactual world and compare the two outcomes. Then, one could argue that the difference in outcomes is really due to the effects of the intervention. As this is impossible, it is necessary to find research designs that succeed in constructing a setting that comes as close as possible to this counterfactual comparison.

In this context, controlled experiments have alluring features that sometimes make them being referred to as the most rigorous of all research designs or even as the ‘gold standard’ (e.g., Angrist 2004). In the basic setting, researchers try to build two groups that are ‘equivalent’ to each other. One group (the treatment group) will be assigned to a specific program, the other group (the control group) will not. Apart from the assignment to the program, the two groups should be the same in terms of all other important characteristics. Like in a counterfactual situation, all outcome differences between the groups could then be attributed to the program. Creating two groups of individuals that are completely equal regarding important aspects such as family background, social status, and the like is not possible as the groups consist of different persons. So it becomes rather a question of probability, and it is important to show that the two groups are equal with sufficiently high probability.

This can be achieved by randomly drawing a sufficiently large number of individuals from a population and randomly assigning them to the treatment and the control group. If consummately implemented, this ensures that the two groups only differ by chance.

2.1.2. An example study: the Perry Preschool Program of early childhood education

We can nicely illustrate the advantages of explicit experiments, but also some limitations and caveats, by an example study that was conducted in the early childhood education sector in the USA. The Perry Preschool Program was launched in 1962 in Ypsilanti, Michigan, when school authorities recognized the low school performance of at-risk children from poor neighborhoods compared with better-off children. In order to improve the social and cognitive development of the children, they decided to test whether an intense intervention of high-quality early childhood education could help these children (see also Barnett 1985).

The crucial feature of this program is that it was conducted as a controlled experiment. In particular, 123 at-risk children of ages three to four were randomly assigned to a treatment and a control group (see Belfield et al. 2006). The 58 children in the treatment group received a high-quality preschool program for one or two (academic) years, including daily center-based care in small groups, home visiting each weekday, and group meetings of the parents. The 65 children in the control group did not receive these services.

One special feature of the study is its extraordinary long-lasting design. The members of both groups were followed up until adulthood and have been evaluated several times. The most recent study was conducted when participants were aged 40. Comparing average social and economic outcomes of the adults in the treatment group and in the control group, the study shows considerably better outcomes for the treatment group in several dimensions. Among other dimensions, on average, the treated individuals had significantly higher earnings, higher probability of high school completion, and fewer incidents of crime than the adults from the control group. The study also reaches a positive net evaluation for the public at large, as the benefits of the program (in terms of higher tax revenues and lower welfare payments) by far exceed the costs. As the assignment to the treatment and the control group was random, it can be concluded that it is the early childhood education that caused the difference in outcomes. The study can thus show a causal effect of the two-year preschool program on the better outcomes.

The study identifies convincing experimental evidence on the effects of the preschool program on later outcomes. However, it is not clear as to what extent the results can be generalized. While the assignment of children to the groups was random, the choice of the underlying sample was explicitly targeted at at-risk children, limiting the 'external validity' of the study to this sample. That is, the causal inference is limited to the specific group of African-American students from disadvantaged backgrounds. If the selection process had been completely random, effects might have differed. But generalized results that are valid for the whole population were not the aim of this specific experiment. Focusing on the effects for specific subgroups is often more interesting and relevant from a political point of view.

2.1.3. Further examples and pertinent issues

Research on several other topics in education has made use of controlled experiments. One of the most well known is the so-called Project STAR in the US state of Tennessee that randomly assigned students and teachers to classes of different sizes. This class size experiment has been subjected to extensive research (e.g., Finn and Achilles 1990) that, among other factors, also takes into account incomplete randomization caused by later switching between classes due to behavioral problems of students and relocation of families (see Krueger 1999). However, substantial worries remain with the implementation of the experiment which may compromise its evidence on the effects of class size on student achievement (see Hanushek 1999).

More recently, several experiments have been implemented to study the effects of incentives on later educational outcomes. A study by Angrist and Lavy (2009) randomly assigns cash incentives to Israeli high school students if they pass the final exam. Bettinger (2008) evaluates the effects of small cash payments to primary school students for successful completion of standardized tests. Both show positive effects of the incentives in the treatment group. Muralidharan and Sundararaman (2009) implement an experiment on teacher performance pay in India. Kremer (2003) summarizes several other randomized evaluations of several educational programs in developing countries, including cash transfers for school attendance, free school meals, free school uniforms, deworming treatments, and increased inputs such as textbooks, flip charts, or a second teacher.

One drawback of explicit experiments is that they tend to be set in a somewhat artificial situation different from real life. Being aware of their participation in an experiment, both individuals of the treatment and the control group could act differently from their 'normal' behavior. Ideally, participants should not be aware of the fact that they are part of an experiment. In their experiment analyzing the use of randomly distributed vouchers for adult training courses, Messer and Wolter (2009) created a setting in which both groups were completely unaware that they were being investigated. The same can be said for situations where random assignment was done for a different reason than an experimental study, as in the case of the random assignment of college roommates which has been used to study peer effects among students (Sacerdote 2001; Zimmerman 2003).[4]

But apart from these specific settings, the fact that participants in an experiment may change their behavior exactly because they are aware that they are being observed (the so-called the Hawthorne effect) is an example of how evidence drawn from controlled experiments can be compromised. This should be especially the case if one result of the experiment appears more favorable to participants than the other. External validity may also be hampered by the fact that a full-scale implementation of a policy could generate general equilibrium effects: If, for example, a small-scale experiment leads to more schooling and, therefore, increased earnings among treated participants, a full-scale intervention might not generate the same effects on earnings because a substantial increase in the supply of highly educated workers may decrease the marginal returns to schooling in general equilibrium. In addition, there can be several factors that complicate both the random draw from the population and the random assignment to the groups. For example, Heckman et al. (2010) show that the Perry Preschool Program did not fully reach random assignment, as treatment and control group individuals were reassigned afterwards.[5] More generally, controlled experiments often suffer from issues of non-perfect implementation, compromising the validity of their results.

When randomization is done explicitly to conduct an experiment in a social setting, people sometimes raise ethical concerns. 'How can those assigned to the control group be denied the blessings of the intervention?', they are inclined to ask. Of course, such objections are only valid if the blessings of the interventions are proven and generally accepted. As in randomized medical trials of new medications, positive effects first have to be convincingly shown before the whole population should be subjected to a treatment. But once the blessings of the intervention have been established, trials are stopped and the medication is released for broader use. In a similar way, although increasing evidence may have made the blessings of early childhood education programs appear obvious today, no such thing could have been said at the start of the Perry Preschool Program, when the relative merits of treatment and non-treatment were hotly debated.

One way to sidestep ethical preoccupations when there is a strong presumption that intervention has positive consequences for participants is to use a set-up of rotating treatments. In such a design, multiple different treatments are allocated to different treatment groups, with each group being exposed to one specific treatment. Although the lack of a control group without any treatment inhibits estimation of the effect of being treated relative to not being treated, such a set-up allows for estimation of the relative effectiveness of the different treatments (which may have been designed to induce the same total costs) by comparing outcomes under one treatment with outcomes under another treatment. It seems to us that such a setting, which is not subject to the ethical objections sometimes raised against controlled experiments, is unwarrantably under-used by educational decision-makers.

2.2. When you cannot serve everyone: lotteries of oversubscribed programs

The second technique covered here is a special case of explicitly controlled experiments that exploits the fact that assignment on oversubscribed programs is often handled by randomized lotteries.

2.2.1. Idea and intuition

Sometimes when an – often charitable – institution aims to implement a specific educational intervention, resources are not enough to finance participation for everybody who is interested. In such a setting, the implementing institution may opt to assign participation by a randomized lottery, so as to give each applicant an even chance for participation.

Among those who apply for the program, this boils down to being an explicitly randomized set-up. If the institution seizes the opportunity of the setting, it aims to observe both the lottery winners and the lottery losers, preferably before but in particular after the implementation of the program. Then, we have randomly assigned treatment and control groups and can estimate the causal effect of the intervention by comparing outcomes between the two.

2.2.2. An example study: voucher lotteries

A well-known example of randomized lotteries of oversubscribed programs is that of programs that offer applicants the chance of winning a voucher that would pay for (part of) the fees required to attend a private school. Examples of such voucher programs that implemented a randomized evaluation include several programs in the USA, namely in New York City, Washington, DC, and Dayton, Ohio. These (privately funded) programs were targeted at low-income families and provided partial funding for private school fees which had to be supplemented by the families. Each program was oversubscribed (i.e., there were more applicants than available vouchers) and participation was decided by lottery. In each case, both those who won a voucher (the treatment group) and those who lost in the lottery and did not receive a voucher (the control group) were observed in terms of background and desired outcomes both before and at several points in time after the program was implemented.

Peterson et al. (2003) use the data from these voucher programs to estimate whether student performance, parental satisfaction, and the educational environment of the schools differed significantly between the treatment and the control group after the intervention. Given random assignment to the groups, such differences could be interpreted as the causal effect of having received a voucher. Among other things, they tend to find that parental satisfaction increased for those who received a voucher and that student achievement on standardized tests increased for those who used it to attend a private school, but only for the subgroup of African-American students. The set-up of the studies allowed them to estimate both the effects of the offer of a voucher (what researchers call the 'intention-to-treat' effect) and the effects of actually using it to switch from a public to a private school (what researchers call the 'treatment-on-the-treated' effect).[6]

2.2.3. Further examples and pertinent issues

In a similar spirit to the previous study, Angrist, Bettinger, and Kremer (2006) estimate the effects of school vouchers in Colombia by exploiting a randomized lottery. A similar approach is also employed by Cullen, Jacob, and Levitt (2006) to estimate the effects of increased choice among specific public schools in Chicago. Bettinger and Slonim (2006) perform a laboratory experiment within the field setting of a voucher lottery in Toledo, Ohio, in order to estimate the effect of school vouchers on altruism.

A particular feature of the setting of randomized lotteries of oversubscribed programs is that the underlying population is just those individuals who applied for participation in the program. This will not necessarily be a random draw from the population at large. In particular, subjects who particularly like the program, who view a particular need for the intervention, or who place particular value on the outcome of the intervention may be more inclined to apply than the average population. As a consequence, results from such studies of oversubscribed programs will be valid estimates of causal effects for those who applied for the program, but their external validity for the population as a whole has to remain an open question.

The most obvious advantage of studies based on randomized lotteries of oversubscribed programs is that they do not require setting up a separate experiment, but rather build on the randomization that is implemented anyway. In addition, these programs tend to refer to field trials that are enacted in a real-world setting, rather than an artificial experimental setting. Similar to explicit experiments, evaluations of randomized lotteries of oversubscribed programs may be subject to the Hawthorne effect and may miss general equilibrium effects. In addition, motivating those who lost in the lottery to participate in subsequent surveys and tests may not be easy.

3. Trying to emulate the dice: natural experiments

The next two techniques aim to exploit variation in observational data that stems from sources that are exogenous to the association of interest. In a sense, they try to mimic the random assignment of controlled experiments by building on incidents where nature or institutional rules and design give rise to random variation.

3.1. Help from outside: instrumental-variable approach

The third technique discussed in this paper aims to identify variation in the exposure to a certain education policy or practice that stems from a particular source that is not correlated with the outcome of interest. This helps to eliminate any part of the variation in treatment that may suffer from endogeneity bias.

3.1.1. Idea and intuition

In the absence of intentional randomization, identifying causal effects is a challenging task. The so-called IV approach is one identification strategy that tries to get close to the set-up of a controlled experiment using observational data. It tries to exploit the fact that nature sometimes makes ‘random assignments’. Therefore, such identification strategies are also referred to as natural experiment or quasi-experiment.

The key idea behind this approach is rather simple. Think of the treatment variable of interest as having two parts: One is subject to the endogeneity problems discussed in the Introduction, for example, because it is correlated with some omitted variable. The other part does not suffer from endogeneity problems and can thus be used for causal identification. The IV approach aims to isolate the latter part of the variation in the treatment variable. This is achieved by using only that part of the variation in the treatment variable that can be attributed to an observed third variable (the instrument) which is not otherwise correlated with the outcome (or with omitted variables that are correlated with the outcome). Having information on such an instrument allows us to isolate variation in treatment that is exogenous to our model and thus to obtain unbiased estimates of the causal effect of treatment on outcome.

The trick of IV estimation is then to find a convincing instrumental variable – one that is associated with the treatment variable (a characteristic called ‘instrument relevance’) but is not correlated with the outcome, apart from the possible indirect effect running through treatment (a characteristic called ‘instrument exogeneity’). If such an instrument can be found, we can identify the treatment effect through a part of the variation in the treatment that is triggered by variation in the instrumental variable, thereby overcoming problems such as reverse causality and omitted variables and achieving consistent estimation.

3.1.2. An example study: the returns to education on the labor market

The study by Harmon and Walker (1995) constitutes a good example that illustrates the use of an IV identification strategy in educational research. In this analysis, the authors estimate the returns to schooling on the labor market, exploiting exogenous changes in the educational attainment of individuals caused by the raising of the minimum school-leaving age in the UK. The validity of standard statistical approaches such as multivariate regressions to estimating the effect of schooling on earnings stands and falls with the assumption that selection into higher and lower levels of schooling is correlated only with observed factors that also correlate with earnings. In the presence of an unobserved factor such as ability which is correlated with schooling and also with earnings, the standard estimates would be biased upward.

To cope with this problem, the study uses variables indicating changes in laws determining the minimum school-leaving age as an instrumental variable for years of schooling. This is possible as individuals in the sample (employed males aged 18–64) faced different minimum school-leaving ages during their youth because two legislative changes raised the minimum school-leaving age from 14 to 15 in 1947 and from 15 to 16 in 1971. The two key assumptions of the IV approach are convincingly met. First, an increase in the minimum school-leaving age induces at least part of the population to stay in school longer. Second, this change in legislation should have no effect on individuals’ earnings other than the indirect effect through increased schooling.

The IV estimates of Harmon and Walker (1995) suggest a rate of returns to schooling of about 15%, that is, each additional year of schooling raises earnings by 15%. This estimate is roughly three times as large as a corresponding standard (OLS) estimate. Moreover, standard tests indicate the presence of endogeneity in the schooling variable in their study which fosters the concern that standard estimates are biased and, hence, IV estimates should be more reliable.

However, one should always interpret IV estimates carefully. If rates of returns differ between individuals (are ‘heterogeneous’), Angrist, Imbens, and Rubin (1996) show that IV procedures estimate the effect of schooling only for that subgroup of the population that complies with the assignment, that is, that actually changes the schooling decision because of a change in the instrument. Therefore, the IV estimate should be interpreted as a so-called Local Average Treatment Effect (LATE), that is, as applying only to the ‘local’ sub-population that is affected by the instrument. In the present case, this suggests that the estimates reflect returns to schooling for those individuals with minimum schooling, whose schooling decisions are affected by the laws on minimum school-leaving ages (although recent evidence by Oreopoulos (2006) suggests that in this specific case, the estimates may be close to the average treatment effect for the population). Thus, effects identified by IV estimation do not necessarily reflect average effects for entire population, raising points of external validity in the same way as true experiments.

3.1.3. Further examples and pertinent issues

Changes in compulsory schooling laws have also been used to examine other important research questions. Black, Devereux, and Salvanes (2008) exploit an increase in minimum schooling from seven to nine years in Norway in the 1960s. They show that increasing mandatory educational attainment through compulsory schooling legislation encourages females to reduce teenage childbearing. Exploiting the same reform, Black, Devereux, and Salvanes (2005) analyze the intergenerational transmission of education. Although simple empirical analyses reveal a positive correlation between educational levels of parents and their children, their analysis suggests that parental education does not causally affect their children's educational attainment. Brunello, Fort, and Weber (2009) make use of various changes in compulsory schooling laws in a cross-country analysis of 12 European countries. They not only confirm that compulsory schooling increases earnings but also show that additional education reduces wage inequality.

Currie and Moretti (2003) use a different source of variation to estimate intergenerational effects of education. They use the availability of colleges in a woman's county in her 17th year as an instrument for maternal education in the USA and find that higher maternal education improves infant health.

Machin and McNally (2007) exploit a change in the rules governing the funding of information and communication technology (ICT) across English school districts in an IV specification to estimate the effect of ICT spending on student performance.

Apart from exogenous variation introduced through changes in policy, researchers have also used exogenous variation literally generated by nature. Haegeland, Raaum, and Salvanes (2008) use variation in educational spending induced by the fact that hydropower plants generate local taxes that finance school expenditures in Norway. Using the location of natural resources as an instrument, they find that standard techniques may be missing the actual effects of school resources on student performance.

Hoxby (2000a) investigates the effects of competition among public schools in the USA, where metropolitan areas differ substantially in the number of separate school districts and thus in choice opportunities. However, the number of school districts itself may be endogenous, for example, because poor performance of existing schools may increase the tendency to open up new districts. To obtain exogenous variation in the extent of school choice, the author uses instrumental variables based on topographics – namely the number of streams – the logic being that students' travel time to school was an important consideration when school district boundaries were initially set in the eighteenth and nineteenth centuries.[7]

When estimating the effect of school choice and competition on student performance in English primary schools, Gibbons, Machin, and Silva (2008) exploit the fact that boundaries of admission districts effectively limit choice and competition. They use the distance from students' homes to the district boundary as an instrument for school choice and the distance from schools to the district boundary as an instrument for school competition.

Using data from the US state of Connecticut, Hoxby (2000b) exploits idiosyncratic variation in cohort sizes to identify the effect of class size on student achievement. This effect is another policy-relevant research question where simple correlations fail to provide reliable guidance for policy-makers because important determinants of class formation are typically unobserved in the data. There is by now an overwhelming evidence that the placement of students into differently sized classrooms between and within schools is severely non-random, being affected by parental choices of residence and school, schools' placement of students into different classrooms within a grade, and school-level placement policies of school systems as a whole. Because natural randomness in the timing of births around school-entry cut-off dates gives rise to variation in cohort sizes which in turn drives exogenous variation in class size across grades in a school, IV specifications using such variation can overcome the endogeneity problems. Woessmann and West (2006) implement an IV strategy similar in spirit to estimate class size effects in a number of countries.

West and Woessmann (2010) exploit the historical pattern that countries with larger shares of Catholics in 1900 tend to have larger shares of privately operated schools even today. Using historical Catholic shares as an instrument for contemporary private competition, they investigate the effect of private competition on student achievement in a cross-country setting.

Ichino and Winter-Ebmer (2004) use variation in school attainment caused by World War II as their source of identification of labor market returns to schooling. They show that Austrian and German individuals who were 10 years old during or immediately after the war went to school for a significantly shorter period than equivalent individuals in others cohort. They exploit this variation in educational attainment to instrument actual schooling.

As these examples illustrate, IV specifications exploit exogenous variation from very different sources, including policy changes, truly natural variations, and historical circumstances. In practice, the key to success in any IV approach is to find a good instrument. In any application, the main assumptions of the IV approach (instrument relevance and exogeneity) must be carefully evaluated. If a convincing instrument is found, causal effects can be well identified even with purely cross-sectional observational data.

The advantage of such quasi-experimental analyses is that they circumvent some of the leading problems with randomized field trials, such as the fact that they are expensive, time-consuming, and difficult to explain to public officials whose cooperation is generally needed. In addition, quasi-experimental studies are not subject to the Hawthorne effect because subjects are not aware that they are part of an experiment, and well-designed natural experiments can capture general equilibrium effects that randomized trials usually cannot.

3.2. When interventions make a jump: regression-discontinuity approach

The fourth technique exploits situations where a treatment sets in discontinuously when a certain assignment variable exceeds a specified threshold level.

3.2.1. Idea and intuition

Another approach in the spirit of natural experiments is the regression-discontinuity (RD) approach. The RD design is used in a specific setting where attendance or non-attendance in a program or intervention is determined by whether a subject falls above or below a certain cut-off value of a specified assignment variable. A standard setting would be where a reform affects only those schools with a certain share of their students having a certain characteristic.

The idea of the RD design then is to compare schools in a sufficiently small range just above and below that cut-off, where those above form the treatment group and those below constitute the control group. The intuition is that these schools will not differ by more than the treatment, because they are very similar in terms of the assignment variable. For example, if schools that have at least 70% students from a disadvantaged background are eligible to a certain intervention, we can compare schools that have 69% and 70% disadvantaged students. They hardly differ in terms of student background, but they do differ in that one school is not eligible for participation and the other one is (see Figure 1).

The comparison of units that are in a sufficiently small range below and above the threshold therefore comes close to an experimental setting with random assignment to treatment and control groups. Any jump or discontinuity in outcomes that can be observed at the threshold can then be interpreted as the causal effect of the intervention. In addition, the fact that the assignment to the treatment and control groups follows a non-linear pattern – the discontinuity at exactly the cut-off value – allows the RD approach to control for any smooth function of the variable determining eligibility. The assumption required for the RD approach to capture the causal effect (the identifying assumption) thus is that there are no other discontinuities around the cut-off.

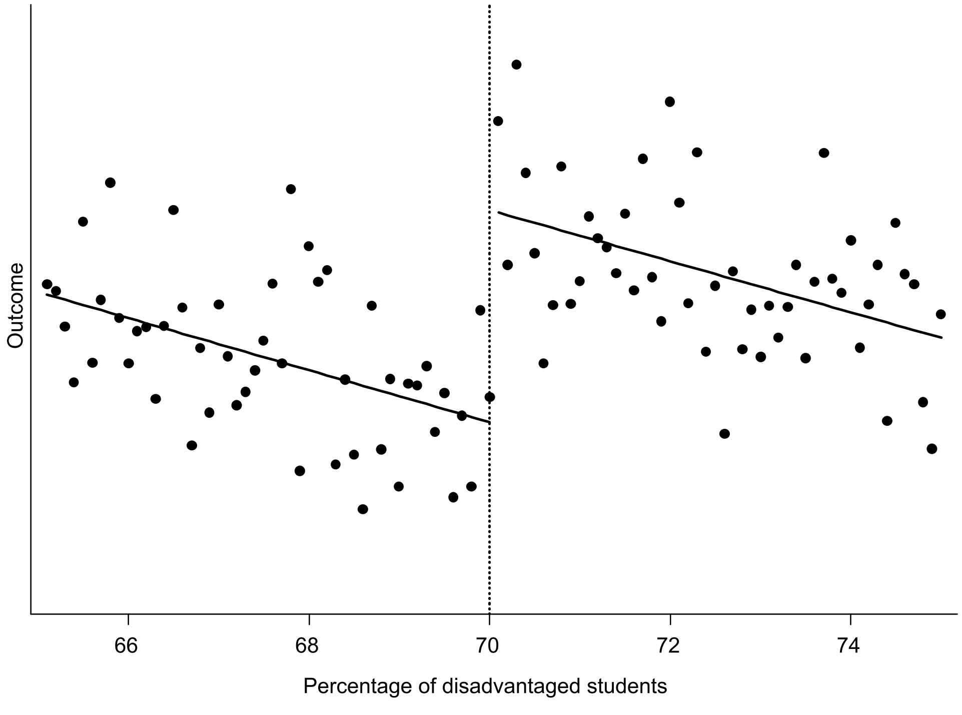

Figure 1. Stylized exposition of the RD design.

3.2.2. An example study: extra funding for disadvantaged students

The study by Leuven et al. (2007) nicely illustrates the use of the RD approach. They evaluate a policy implemented in the Netherlands in 2000 and 2001. The government decided to provide extra funding to schools with a high share of disadvantaged students who belong to ethnic minorities or whose parents have a low level of education. The criterion for being selected for the subsidies was a sharp discontinuity: Schools with a share of at least 70% of their students having a disadvantaged background were eligible to obtain the extra funding.

There were two programs in the two years, one consisting of an extra payment per teacher and the other of a one-time payment per student that was earmarked for computers and software. The authors can use data on students' performance in several standardized tests before and after the introduction of the program (both from schools that qualified for the subsidy and from schools that did not) to study the effects of the intervention. A nice feature of the RD design is that it has a straightforward graphical depiction: When plotting the outcome of interest against the assignment variable (the share of disadvantaged students in this case), there should be a clear jump in the outcome at the assignment threshold that determines the treatment status of the schools. In the Dutch study, there is no such jump in test scores, or – if anything – it is negative.

In order to show whether this downward jump is really caused by the subsidies, they restrict their sample to schools that have a share of disadvantaged students of ±5 (or ±10) percentage points around the threshold of 70%. Then, they compare the difference in the average test scores between the schools above and below the threshold, controlling for any smooth effect that the share of disadvantaged students may have on average test scores. Their results corroborate the graphical finding that average test scores of the treated schools are, if anything, worse than those of the non-treated schools. As the assignment to treatment and control group close to the threshold is supposed to be random, in most specifications, they identify no significant causal effect of the subsidies on students' achievement. For some outcomes – in particular, the effect of extra funding for computers and software on girls' achievement – they even find a significant negative effect of the subsidy, which may indicate that computer-aided instruction is not the most effective way of teaching in this circumstance.

The study by Leuven et al. (2007) is also able to address several potential caveats of the RD approach. In principle, schools could manipulate the share of disadvantaged students if they were able to anticipate the intervention and the eligibility rule, which would undermine identification of the causal effect. However, because the date to which the cut-off rule refers is long before the announcement of the subsidies, this is unlikely in the given study. Moreover, there are only few schools that do not comply with the 70% rule in this study, giving rise to a reasonably sharp change of treatment status at the cut-off point.

3.2.3. Further examples and pertinent issues

Several studies in the literature on the effects of class size on educational outcomes exploit discontinuities due to maximum class size rules: Class size drops discontinuously whenever grade enrolment would lead to class sizes that exceed the maximum size determined by a specific rule. One can then compare classes just above and below these thresholds to obtain exogenous variation in class size (see Angrist and Lavy 1999 for Israel; Woessmann 2005 for several European countries; and Leuven, Oosterbeek, and Rønning 2008 for Norway).

Another application of the RD design uses specified school entry cut-off dates that lead to the effect that school entry ages vary due to the month of birth of the children. Children born just after the respective cut-off date are nearly one year older when entering school than children born just before the cut-off date (see Bedard and Dhuey 2006; Puhani and Weber 2007). Studies exploiting maximum class size rules or school entry cut-off dates are classical examples for the so-called fuzzy RD designs. In contrast to the so-called sharp RD designs, where the cut-off unequivocally divides observations into a treatment and a control group, in fuzzy RD designs, not all observations comply with the rule. In the school entry example, some children born just after the cut-off date nevertheless enter school earlier than they are supposed to and vice versa. In fuzzy RD approaches, treatment and control observations are thus observed both below and above the cut-off. Rather than exploiting sharp changes of treatment status at the cut-off point, fuzzy RD designs exploit discontinuities in the probability of treatment. They can thus be interpreted as an IV approach where the discontinuity acts as the instrument for treatment status.

Ludwig and Miller (2007) exploit a discontinuity in the technical assistance that an office of the US federal government provided to the 300 poorest counties to develop proposals to participate in the national preschool program called Head Start. Given that this assistance discontinuity generated a large and lasting discontinuity in Head Start funding rates, the authors can estimate the effects of Head Start on children's health and educational attainment.

Garibaldi et al. (2007) study the effect of tuition fees on the probability of late graduation from university, exploiting the fact that tuition fees at Bocconi University in Milan are subject to discontinuous changes with respect to family income. Gibbons and Machin (2006) use discontinuities in school admissions patterns caused by admission district boundaries to estimate whether school quality affects housing prices. Lavy (2009) exploits a discontinuity in the assignment of schools to a program of performance-related teacher pay in Israel to estimate its effect on student outcomes. Lavy (forthcoming) uses a geographical RD approach that compares students in an area (Tel Aviv) that enacted free school choice to students in neighboring areas that were not subject to the treatment. West and Peterson (2006) exploit discontinuities in a school accountability program in Florida at a specific performance grade to estimate whether the threat to very poorly performing public schools that their students obtain a voucher to attend private schools affects public schools' performance.

The examples illustrate that there is a rich number of cases where educational policies and practices are implemented in a manner that involves a discontinuity which allows for evaluation through the RD approach. A problem with implementing the RD design is that it is not always possible to find enough observation units in an area just below and above the respective cut-offs. One solution is to increase the bandwidths around the thresholds, but this reduces the probability that the units above and below the threshold only differ by their treatment status. Moreover, due to the local identification around the threshold, external validity is also an issue for estimates based on the RD approach.

4. 'Fixing' the unobserved: methods using panel data

The remaining two techniques aim to account for endogeneity in observational data by exploiting variation where subjects are observed several times, usually (but not necessarily) at several points in time. Such two-dimensional datasets are called panel data. The common hope of all panel data-based techniques is to be able to control for intervening factors even though they might be unobserved in the data. This is possible as long as these unobserved factors are 'fixed' along the second dimension of the dataset (e.g., constant over time).

4.1. Does the difference differ? Differences-in-differences approach

The fifth technique discussed in this paper builds on datasets that observe each subject at least twice and where part of the subjects changes its treatment status between the two incidents of observation. The incidents usually refer to two points in time, but they may also refer to two other dimensions such as different grade levels or subjects.

4.1.1. Idea and intuition

Differences-in-differences (DiD) estimation is a simple panel data method applied in situations when certain groups are exposed to a treatment and others are not. Consider the simple case of two groups and two periods. In the first period, none of the groups is exposed to treatment. In the second period, only one of the groups gets exposed to treatment, but not the other. As a hypothetical example, think of two classes in a given school observed at the beginning and the end of a school year. During this school year, only students in one of these two classes have additional afternoon lessons. DiD estimation can then be used to answer the question, 'What is the effect of additional lessons in the afternoon on student achievement?' The DiD identification strategy now consists of taking two differences between group means in the following way (illustrated in Figure 2). First, we compute the difference in the mean of the outcome variable between the two periods for each of the groups; this is the first difference. In the hypothetical example, the first difference simply corresponds to the change in average test scores for each group between the beginning and the end of the school year. Then, we take the second difference – between the differences calculated for the

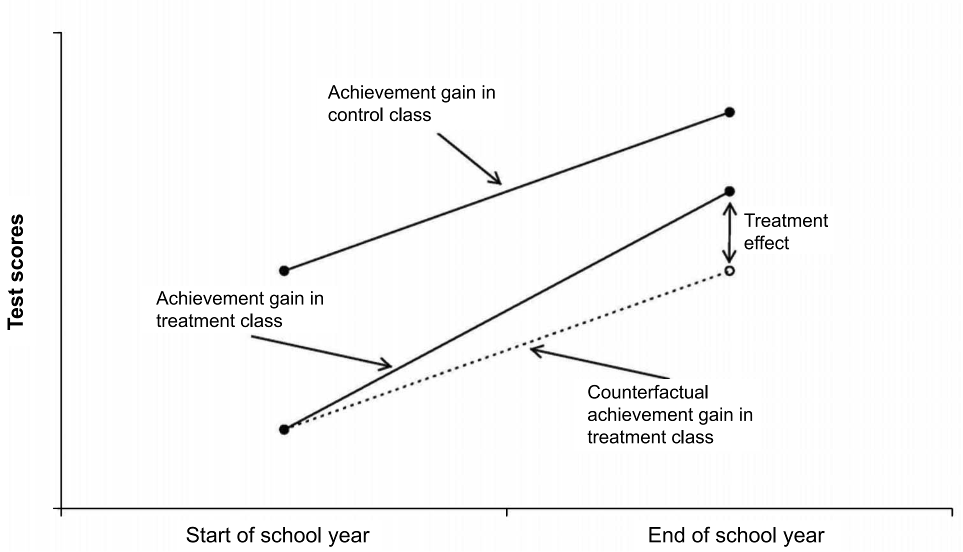
Test scores
Achievement gain in control class
Treatment effect
Achievement gain in treatment class
Counterfactual achievement gain in treatment class
Start of school year
End of school year

Figure 2. Stylized exposition of identification in the DiD model.

two groups in the first stage (which is why the DiD method is sometimes also labeled 'double differencing' strategy). This second difference measures how the change in outcome differs between the two groups, which is interpreted as the causal effect of the causing variable. Hence, in our example, the effect of afternoon lessons on student learning is identified by comparing the gains in average test scores over the school year between the two classes.

The idea behind this strategy is simple: The two groups might be observationally different. That is, the group-specific means might differ in the absence of treatment. However, as long as this difference is constant over time (in the absence of treatment), it can be differenced out by deducting group-specific means of the outcome of interest. The remaining difference between these group-specific differences must then reflect the causal effect of interest.

The key assumption required for this approach to identify the causal effect is that the group-specific trends in the outcome of interest would be identical in the absence of treatment. In terms of the hypothetical example, the identifying assumption is that both classes would have experienced the same increase in test scores over the school year in the absence of afternoon lessons. The assumption that the treatment class would have experienced a counterfactual achievement gain identical to the observed achievement gain in the control class is illustrated by the dotted line in Figure 2. The plausibility of this identifying assumption depends on the specific setting to which DiD estimation is applied. In any case, the identifying assumption of the DiD approach is less restrictive than the assumption implicitly made in standard classical methods, namely that the two groups are identical in terms of all relevant unobserved factors.

4.1.2. An example study: a reform of compulsory schooling and tracking

The study of Meghir and Palme (2005) constitutes a good real-world example of a DiD research design. In this study, the authors evaluate the effects of a Swedish educational reform designed in the late 1940s that increased compulsory schooling, imposed a nationally unified curriculum, and abolished school tracking. DiD estimation is feasible because implementation of the reform varied between municipalities. The study focuses on individuals from two birth cohorts, 1948 and 1953. For a substantial part of Swedish municipalities (which we will call the 'switching municipalities'), these cohorts were assigned to different school systems: the 1948 cohort to the old system and the 1953 system to the new system. At the same time, in some municipalities, both cohorts were assigned to the old system, while in other municipalities, both cohorts were assigned to the new system.

The DiD identification rests on the comparison of the change in average outcomes between the 1948 and the 1953 cohorts in the switching municipalities with the change in average outcomes between the same cohorts in the municipalities that did not change with respect to the new system. This empirical strategy implicitly makes the identifying assumption that in the absence of reform, the changes in the average outcomes between the 1948 and the 1953 birth cohorts living in the municipalities that adopted the reform would have been the same as the changes for those living in the municipalities whose reform status remained the same for these two cohorts (conditional on other observed characteristics).

The study finds that the reform improved both average educational attainment and earnings of the Swedish population, and in particular of children from less educationally advantaged family backgrounds. This analysis nicely illustrates that DiD strategies are often well suited to estimate the causal effect of sharp changes in education policies or practices, providing policy-makers with vital information even in the absence of controlled or natural experiments.

4.1.3. Further examples and pertinent issues

DiD estimation of a policy change is a widely used identification strategy. Pekkarinen, Uusitalo, and Kerr (2009) investigate the relationship between features of the school system and the intergenerational persistence of earnings. Their empirical strategy exploits a Finnish comprehensive school reform enacted between 1972 and 1977 that shifted the age at which children are tracked into differing-ability secondary schools from 10 to 16 and imposed a uniform academic curriculum on entire cohorts until the end of lower secondary school. As in Meghir and Palme (2005), DiD estimation is feasible because the reform was implemented gradually across municipalities during a six-year period. In both cases, it has to be assumed that the fact that some municipalities enacted the reform earlier than others is not correlated with unobserved characteristics of the municipalities that are themselves correlated with the subsequent change in outcomes. If municipalities that would anyways have embarked on a different outcome trajectory than the others are more inclined to enact the reform first, then the DiD approach will not yield an unbiased estimate of the causal reform effect.

Dynarski (2003) estimates the effect of student aid on college attendance and completion. In contrast to the previously discussed studies, this analysis uses the elimination of a student aid program for identification. In the USA before 1982, 18- to 22-year-old children of deceased, disabled, or retired social security beneficiaries received monthly payments while enrolled full time in college. This program was terminated in 1982. The DiD strategy rests on the comparison of college attendance and completion for eligible students before and after the elimination of the program compared with the corresponding difference for students who were never eligible for this student aid.

Another study by Jacob (2005) investigates whether test-based accountability raises student performance. More precisely, this study examines the impact of an accountability policy implemented in public schools in Chicago in 1996–1997. DiD estimation is applied here in two ways to identify the effect of the reform. The first is to compare changes in test scores for individual students before and after the reform in Chicago public schools. The second is to compare the difference in test scores before and after the reform in the school district of Chicago with the corresponding change in other similar school districts that did not implement such a reform. In a similar study, Hanushek and Raymond (2004) use a DiD set-up to analyze how the differential introduction of accountability across US states affected the relative change in student performance.

Bénabou, Kramarz, and Prost (2009) use the DiD approach to analyze the effects of the policy of education priority zones in France, which channels additional resources to disadvantaged schools. They exploit the implementation of the policy in different schools at different points in time. Machin and McNally (2008) examine the introduction of the literacy hour, a specific teaching innovation introduced in English primary schools in the 1990s. Their DiD strategy is based on a pre- and post-introduction comparison of schools that did or did not introduce the literacy hour.

However, a DiD strategy does not always have to include a time dimension. For example, Hanushek and Woessmann (2006) investigate whether educational tracking affects school performance and inequality in the school system. Their research design is based on comparing outcome differences between primary school (when no country uses tracking) and secondary school (when some countries use tracking and others do not) in countries with and without tracked school systems. Jürges, Schneider, and Büchel (2005) use a DiD strategy to identify the causal effect of central exams on student performance. Their double-differencing strategy exploits the fact that in some German states, mathematics is tested in central exams while science is never tested in central exams. The first difference is thus given by the difference in student test scores between subjects.

4.2. Information overflow: panel data techniques and taking out fixed effects

The sixth group of techniques generalizes the previous approach and requires rich datasets that allow accounting for unobserved individual fixed effects.

4.2.1. Idea and intuition

In recent years, extensive datasets have become available, in particular in a few US states, which provide performance data for every student in the public school system and allow following up each student over several years.[8] Such extensive panel datasets allow an even more extensive treatment of cases where unobserved differences are constant over time than traditional DiD approaches. As discussed in the Introduction, neglecting such unobserved heterogeneity leads to results that cannot disentangle the effect of the policy intervention from the influence of other unobserved (or imprecisely observed) factors. Moreover, these more extensive panel data techniques can be applied to the case where unobserved factors are heterogeneous across individuals (rather than groups).

One of the most prominent factors in educational research that is usually not observed by researchers is the ability of individuals. For many problems that empirical educational research deals with, a possible ‘ability bias’ is a typical challenge. For example, if we try to evaluate whether the attendance of private schools improves students’ achievements, our estimates are only reliable if we can exclude that they are driven by differences in ability between students that choose to attend private and public schools. In the absence of a clear measure for this factor and of a controlled or natural experiment, researchers sometimes try to circumvent this drawback by observing adequate proxy variables such as family background, parental education, and the like. However, such an approach will generally not be able to solve the central problem of unobserved individual heterogeneity.

As long as the factor that leads to unobserved heterogeneity – such as individual ability – is constant over time, it can be fixed by ignoring any variation in the level of outcomes across individuals and only focusing on changes over time. For this, we need panel data that allow us to observe the same individual at several points in time. In such a setting, we can control for fixed effects of each individual in that we account for an indicator variable that takes out mean differences between individuals so that only changes over time in inputs and outcomes are used to identify the effects of interest. This way, the estimation is able to control for unobserved but fixed heterogeneity across units of observation. In the most elaborate settings, such fixed effects can be introduced at the level of students, teachers, and schools.

A special case of such models is the classical value-added approach that controls for previous outcomes of the same unit when estimating the effect of some input or intervention on current outcomes. In the special case of observing the same outcome (e.g., a test score) for a unit at two points in time, this approach estimates the effect of a specific education policy conditional on the previous outcome. This value-added approach is only a special case of more extensive models that use cross-sectional information of the same units (individuals, classes, or schools) at several points in time in order to take out the effects of any time-invariant factor, or fixed effect (see Todd and Wolpin 2003 for additional discussion).

Using the time dimension is only one approach to take out fixed effects. Even without having observations for the same unit at several points in time, one can exclude fixed effects by using special datasets. Specific studies are able to use information on contemporaneously observed academic achievements in two subjects of the same individual, which allows them to take out fixed effects of a student that affect all subjects alike. Other approaches use data on siblings or twins to address unobserved heterogeneity. The advantages and caveats of these special approaches will be discussed below.

4.2.2. An example study: peer effects in school

The study by Vigdor and Nechyba (2007) which evaluates the effects of peer characteristics on academic achievement is a nice example of the use of fixed-effects models to identify causal effects, which also highlights the limitations of restricted models. Peer effects are widely discussed in the education literature as a potentially important determinant of students' outcomes. The literature on peer effects deals with several identification problems. First of all, it is difficult to disentangle the effects of the peer group on an individual's outcomes from the effects of the individual on the group (see Manski 1993). Moreover, the choice of a peer group of an individual often is not exogenous, meaning that certain individuals, for example well-performing students, select themselves into specific peer groups that are already made up of other well-performing students. The authors analyze the effects of the characteristics of the student body in schools on academic achievement of individuals. Being able to use longitudinal data from all students in North Carolina public schools, they try to estimate whether the average achievement level of a student's peers affects the performance of the individual student.

Their data allow them to circumvent one crucial identification problem of peer effects, namely ability sorting across different schools. Certain schools could mainly attract high-ability students, others predominantly low performers. The effect of the peer group could then not be distinguished from the characteristics of the student body attracted by a specific school. As the authors can match students with their classmates, they can use school fixed effects and eliminate the bias arising from systematic differences between schools. By estimating the effects of peer ability, measured by standardized test scores of a student's peers in the third grade, on student achievement in the fifth grade and using school fixed effects, they find a significant positive association between average peer performance and student achievement.

Nevertheless, identification of causal effects remains a challenge if sorting takes place not only across schools but also within schools across classrooms. Parents who are especially keen on high-quality education for their children could achieve to have their children in classes with high-ability peers and/or match them with high-quality teachers. So the estimated association between peer and individual performance could also reflect the fact that similar-ability children tend to attend the same class and/or have the same teacher. Observing several cohorts of fifth-grade students, the authors can – in addition to accounting for school fixed effects – match students with their teachers and include teacher fixed effects to avoid distortions arising due to different time-invariant teacher quality across classrooms. After accounting for teacher fixed effects, any significant estimate on peers' performance vanishes. This result suggests that even in a specification that includes school fixed effects, the association between peer and individual performance reflects a mere correlation rather than a causal effect of peer performance on individual performance. In reality, causal peer effects may be much smaller (or even zero) than more restricted techniques may suggest.

This result is corroborated by another test enabled by the panel nature of the data. The authors find that a significant association is evident between the performance of an individual's peers in the fifth grade and her own performance in the fourth grade, even after controlling for the characteristics of peers in the fourth grade. Such a result that the appearance of apparent peer effects predates the actual exposure to the specific peer group is a tell-tale sign that the estimated association actually reflects selection into peer groups, rather than the causal effect of peer groups on individual performance.

These results reveal how panel data techniques that control for fixed effects can be put to informative use, while alerting us to the limitations of fixing effects at too aggregate a level. In this case, unobserved, time-invariant heterogeneity is not eliminated when controlling for school fixed effects, because additionally controlling for teacher fixed effects completely changes the results.[9]

4.2.3. Further examples and pertinent issues

One recent application of panel data techniques that control for student fixed effects is the estimation of teacher effects on student learning. As variation in teacher quality is difficult to measure, recent studies have started to use matched student–teacher data, where teacher influence can be observed by estimating fixed effects (see Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Clotfelter, Ladd, and Vigdor 2007). Such applications are able to control for individual student fixed effects by observing the same student more than once, thereby focusing on differences in the performance of the same student with different teachers. Broadly speaking, teacher fixed effects in such applications can serve as measures of the overall effect of teachers on student outcomes, which can then be associated with observed teacher characteristics. In a similar way, Hanushek et al. (2003, 2007) control for unobserved student heterogeneity by including student and school fixed effects in rich panel datasets to estimate the effects of peer ability and of charter schools on student achievement.

Other studies apply panel data techniques to longitudinal datasets that do not provide such rich student information. Bauer (2002) uses individual fixed effects to account for unobserved heterogeneity when estimating the labor market effects of the so-called over- and undereducation of workers, suggesting that standard methods lead to highly misleading results on this topic. Böhlmark and Lindahl (2008) use Swedish panel data to take out municipality fixed effects when estimating the effects of a voucher reform on educational achievement. Bonesrønning, Falch, and Strøm (2005) use school fixed effects to estimate the effect of the composition of the student body, such as the share of students from minority groups, on teacher supply and demand in Norway. They also combine fixed effects with the IV approach using the arrival of refugees from former Yugoslavia as an instrument for the minority share. Falch (2008) uses school fixed effects to estimate how wages affect the supply of teachers at the school level under a specific wage-setting regime in Norway.

In the absence of rich repeated information on educational outcomes, several recent studies have made use of datasets that observe the performance of the same individuals in different subjects. By including student fixed effects, these studies can account for any type of unobserved student characteristics that affects different subjects in the same way and estimate the effect of teacher characteristics or teaching practices not across but within students (see Dee 2005, 2007; Ammermüller and Dolton 2006; Schwerdt and Wuppermann 2009). If teachers are also observed in terms of characteristics that differ across subjects, such as their subject knowledge, fixed teacher effects can even be controlled for in addition to fixed student effects when estimating the effects of subject-variant teacher characteristics (Metzler and Woessmann 2009).

Another strategy that tries to overcome unobserved heterogeneity by controlling for fixed effects is the use of data on siblings and twins. When observing different siblings or twins from the same family, the analysis can control for fixed effects for the specific family background, or with identical twins, even for genetic differences. Several studies have applied siblings or twins models for identification of the effect of schooling on earnings. Data on siblings or twins are supposed to reduce the endogeneity problem, as siblings or twins are presumably more alike in terms of unobserved factors such as ability and parental emphasis on education than a random pair of individuals. Griliches (1979) and Card (1999) provide surveys of twin/sibling-based estimates of returns to schooling. Behrman and Rosenzweig (2002) use a sample of identical twins to estimate the effect of parents' education on the education of their children. Garces, Thomas, and Currie (2002) use comparisons among siblings to estimate the effect of participation in the US Head Start preschool program on school completion, higher education, earnings, and crime. Black, Devereux, and Salvanes (2007) apply a twin model to estimate the effect of birth weight on education in Norway.

While twin studies are suitable for tackling many omitted variable problems, further issues may limit identification. A fundamental challenge in twin studies is whether it is convincing to view the source of the variation as exogenous: If two supposedly identical twins with supposedly the same environment receive different amounts of education, it seems likely that in the end they did differ in certain dimensions unobserved by the researcher. Furthermore, the inclusion of fixed effect tends to aggravate the problem of measurement error in the explanatory variable, biasing effects toward zero in twin studies (see Ashenfelter and Krueger 1994). Finally, the extent to which results based on twins, who grow up in a specific situation, generalize to the whole population is unclear, raising issues of external validity.

Despite the substantial advancements made over recent years, it remains a fact that panel data techniques can account for unobserved heterogeneity only inasmuch as the relevant unobserved characteristics do not differ systematically over the panel dimension. Unobserved interferences that are not constant over time (or subjects or twins) cannot be fixed by standard panel data techniques alone and hence require the experimental methods discussed above. Consider student motivation as an example of an unobserved factor that is positively correlated with student outcomes. Panel data models will only be able to fix bias from differential motivation if it does not change over time. But students' motivation can change over time, and such time-varying unobserved heterogeneity will bias panel data results if the change is correlated with the intervention of interest for other reasons than being its consequence. Furthermore, panel data techniques can only solve the endogeneity problem if changes to the variable of interest are genuine changes (not due to measurement error) and exogenous.

5. Conclusion: the need for more policy evaluation

This paper aims to familiarize government officials, civil servants, educational managers, and other practitioners, as well as researchers from outside the econometric field, with state-of-the-art evaluation methods that can produce causal evidence suitable for guiding policy in the field of education. If policies are meant to improve on outcomes, they should not be based on ideology or wishful thinking but rather on proven effectiveness. But obtaining convincing evidence on the effects on specific education policies and practices is not an easy task. As a precondition, relevant data on possible outcomes has to be gathered. What is more, showing a mere correlation between a specific policy or practice and potential outcomes is no proof that the policy or practice caused the outcome. For policy purposes, mere correlations are irrelevant, and only causation is important. What policy-makers care about is what would really happen if they implemented a specific policy or practice – would it really change any outcome that society cares about? In order to implement evidence-based policy, policy-makers require answers to such causal questions.

This paper has surveyed new techniques that researchers have designed to go from correlation to causation. By avoiding technical language as far as possible, it is hoped that this non-technical guide proves useful to education policy-makers, practitioners, and their advisors in their task of handling the problems and possible solutions to the causal evaluation of education policies and practices. Ultimately, the different methods aim to implement an 'experimental' design where the population is randomly divided into a treatment group that is subjected to the policy and a control group that is not. Thereby, they overcome the problem that unobserved confounding factors may give rise to spurious associations and allow for a causal evaluation of the policy.

By no means do we want to claim that the techniques described here are the only ones by which causal evaluation is possible. We also hasten to refer to the discussed underlying assumptions that have to be met in order for the different methods to yield causal estimates. But the methods discussed in this paper help to point out how more standard research designs can go wrong and yield results that do not capture causal effects. Also, the insight that experimental and quasi-experimental studies are best positioned to discern causal effects does not invalidate studies using more traditional techniques such as standard multivariate regressions. In many circumstances where experimental designs cannot yet be implemented, more traditional studies may serve to provide first hints of where to look and what to expect. But the issues and findings discussed in this paper caution about possible limits to the causal interpretation of such studies.

For a better understanding of the causal relationships of interest, there is a strong need for proper and regular evaluation of each educational reform. Convincing evaluation should be a central part of the design of any education policy and practice. This way, researchers, policy-makers, and practitioners alike could learn what really works and what does not in education policy and practice. Results of such evaluations could then be combined with cost information to find out the most beneficial and cost-effective way to achieve desired outcomes. While such information is necessary for effective education policy and practice, of course, it is not sufficient. Evidence alone does not lead to policy chance, not least because certain political and vested interests often stand in the way. Such considerations of how to get from evidence to action go beyond the scope of this paper. Still, sound evaluation is an inevitable prerequisite to ensure that future policies and practices will be more successful than previous ones in achieving the outcomes that they strive for.

In order to achieve this, adequate data have to be available. There is need to accompany any reform of education policy and practice with a properly designed evaluation technique and collect the data required to perform the evaluation. More often than not, the data required for convincing evaluation will not be available off the shelf but will have to be collected as part of the reform. Also, the proper evaluation techniques for any education reform will usually not be waiting to be simply lifted off the shelf but instead will need to be carefully designed with respect to the specific problems that arise with the particular program that is to be evaluated. In order to set up convincing evaluation designs, it may also prove helpful to involve evaluators in designing the policy ex ante. With regard to data collection and evaluation design, especially Europe has to catch up in order to learn what works and what does not in education policy and practice.

In medical research, experimental evaluation techniques are a well-accepted standard device to learn what works and what does not. No one would treat large numbers of people with a certain medication unless it has been shown to work. Experimental and quasi-experimental studies are the best way to reach such an assessment. It is hoped that a similar comprehension is reached in education so that future education policies and practices will be able to better serve the students.

Acknowledgements

This paper is an adapted version of an analytical report prepared by the European Expert Network on Economics of Education (EENEE) for the European Commission. We would like to thank Lucie Davoine, the fellow EENEE members (see www.education-economics.org), in particular Giorgio Brunello, Torberg Falch, Daniel Munich, and George Psacharopoulos, seminar participants at the Ifo Institute in Munich, the NESSE members Ides Nicaise, Kathleen Lynch, and Susan Robertson, and two anonymous referees for comments and suggestions that improved the exposition of the paper.

Notes

[1] For more technical treatments, the interested reader is referred to the technical literature on the subject such as Angrist and Krueger (1999), Todd and Wolpin (2003), and Webbink (2005). A standard introductory textbook to the econometric techniques that often uses examples drawn from the education field is Stock and Watson (2006). Similarly, Angrist and Pischke (2009) provide a guide for econometric practitioners that also contains examples from education.

[2] Other possible sources of endogeneity include self-selection (objects with different characteristics can choose whether to be treated or not) and simultaneity (treatment and outcome are choice variables that are jointly determined). In econometric terms, measurement error in the treatment variable can also be interpreted as an endogeneity problem, because it gives rise to a particular form of association between treatment and outcome (one that generally biases the estimates toward finding no effect, even if there was one).

3. Matching techniques can still improve on the formation of proper treatment and control groups when they are combined with one of the approaches discussed below, where the latter ensures exogeneity of treatment and thus causal interpretation. For examples of such combinations from the education area, see, for example, Lavy (2009) and Machin and McNally (2008). Galindo-Rueda and Vignoles (2007) is an example of applying matching techniques to the estimation of the effects of comprehensive versus selective schooling; evidence provided in Manning and Pischke (2006) suggests that matching-type estimation alone does not solve the endogeneity problem inherent in this setting.
4. Ideally, researchers conducting the experiment should also be unaware whether participants are part of the treatment or the control group during the experiment. This 'double-blind' assumption is particularly important when experimenters are involved during the experimental phase or personally collect information on the outcome of interest, like physicians in medical trials.
5. Among other things, some individuals originally assigned to treatment who had working mothers were swapped with control individuals whose mothers were not employed, because it was deemed difficult for working mothers to participate in home visits assigned to the treatment group.
6. See Krueger and Zhu (2004) and Peterson and Howell (2004) for additional discussion of the results of the New York City voucher lottery.
7. See Rothstein (2007) and Hoxby (2007) for further discussion of this approach.
8. While such extensive datasets are not (yet) available in European countries, data in some European countries – such as extensive register data in several Scandinavian countries and a new student performance database in the UK – come close.
9. In an additional step, the authors go beyond fixed-effects specifications to exploit a special feature of the North Carolina school system in order to use a plausibly exogenous shock to peer group composition attributable to the opening of new schools in an IV specification, corroborating the result that there are no significant peer effects.

References

Ammermüller, A., and P. Dolton. 2006. Pupil–teacher gender interaction effects on scholastic outcomes in England and the USA. ZEW Discussion Paper 06-060. Mannheim, Germany: Centre for European Economic Research.

Angrist, J.D. 2004. American education research changes tack. *Oxford Review of Economic Policy* 20, no. 2: 198–212.

Angrist, J.D., E. Bettinger, and M. Kremer. 2006. Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review* 96, no. 3: 847–62.

Angrist, J.D., G.W. Imbens, and D.B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91, no. 434: 444–55.

Angrist, J.D., and A.B. Krueger. 1999. Empirical strategies in labor economics. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card, vol. 3A, 1277–366. Amsterdam: North-Holland.

Angrist, J.D., and V. Lavy. 1999. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114, no. 2: 533–75.

Angrist, J.D., and V. Lavy. 2009. The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review* 99, no. 4: 1384–414.

Angrist, J.D., and J.-S. Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Ashenfelter, O., and A. Krueger. 1994. Estimates of the economic return to schooling from a new sample of twins. *American Economic Review* 84, no. 5: 1157–73.

Barnett, S. 1985. Benefit-cost analysis of the Perry Preschool Program and its policy implications. *Educational Evaluation and Policy Analysis* 7, no. 4: 333–42.

Bauer, T. 2002. Educational mismatch and wages: A panel analysis. *Economics of Education Review* 21, no. 3: 221–9.

Becker, S.O., and A. Ichino. 2002. Estimation of average treatment effects based on propensity scores. *Stata Journal* 2, no. 4: 358–77.

Bedard, K., and E. Dhuey. 2006. The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics* 121, no. 4: 1437–72.

Behrman, J.R., and M.R. Rosenzweig. 2002. Does increasing women's schooling raise the schooling of the next generation? *American Economic Review* 92, no. 1: 323–34.

Belfield, C.R., M. Nores, S. Barnett, and L. Schweinhart. 2006. The High/Scope Perry Preschool Program. *Journal of Human Resources* 41, no. 1: 162–90.

Bénabou, R., F. Kramarz, and C. Prost. 2009. The French zones d'education prioritaire: Much ado about nothing? *Economics of Education Review* 28, no. 3: 345–6.

Bettinger, E. 2008. Paying to learn: The effect of financial incentives on elementary school test scores. Paper presented at the CESifo/PEPG Conference 'Economic Incentives: Do They Work in Education?', May 16–17, in Munich, Germany.

Bettinger, E., and R. Slonim. 2006. Using experimental economics to measure the effects of a natural educational experiment on altruism. *Journal of Public Economics* 90, nos. 8–9: 1625–48.

Black, S.E., P.J. Devereux, and K.G. Salvanes. 2005. Why the apple doesn't fall far: Understanding intergenerational transmission of human capital. *American Economic Review* 95, no. 1: 437–49.

Black, S.E., P.J. Devereux, and K.G. Salvanes. 2007. From the cradle to the labor market? The effect of birth weight on adult outcomes. *Quarterly Journal of Economics* 122, no. 1: 409–39.

Black, S.E., P.J. Devereux, and K.G. Salvanes. 2008. Staying in the classroom and out of the maternity ward? The effect of compulsory schooling laws on teenage births. *Economic Journal* 118, no. 530: 1025–54.

Böhlmark, A., and M. Lindahl. 2008. Does school privatization improve educational achievement? Evidence from Sweden's Voucher Reform. IZA Discussion Paper 3691. Bonn: Institute for the Study of Labor.

Bonesrønning, H., T. Falch, and B. Strøm. 2005. Teacher sorting, teacher quality, and student composition. *European Economic Review* 49, no. 2: 457–83.

Brunello, G., M. Fort, and G. Weber. 2009. Changes in compulsory schooling, education and the distribution of wages in Europe. *Economic Journal* 119, no. 536: 516–39.

Card, D. 1999. The causal effect of education on earnings. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card, vol. 3A, 1801–63. Amsterdam: North-Holland.

Clotfelter, C.T., H.F. Ladd, and J.L. Vigdor. 2007. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review* 26, no. 6: 673–82.

Commission of the European Communities. 2007. Towards more knowledge-based policy and practice in education and training. Commission Staff Working Document SEC 1098. Brussels: Commission of the European Communities. http://ec.europa.eu/education/policies/2010/doc/sec1098_en.pdf

Cullen, J.B., B.A. Jacob, and S. Levitt. 2006. The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica* 74, no. 5: 1191–230.

Currie, J., and E. Moretti. 2003. Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *Quarterly Journal of Economics* 118, no. 4: 1495–532.

Dee, T.S. 2005. A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review* 95, no. 2: 158–65.

Dee, T.S. 2007. Teachers and the gender gaps in student achievement. *Journal of Human Resources* 42, no. 3: 528–54.

Dynarski, S.M. 2003. Does aid matter? Measuring the effect of student aid on college attendance and completion. *American Economic Review* 93, no. 1: 279–88.

Falch, T. 2008. The elasticity of labor supply at the establishment level. Industrial Relations Section Working Paper 536. Princeton, NJ: Princeton University.

Finn, J.D., and C.M. Achilles. 1990. Answers and questions about class size: A statewide experiment. *American Educational Research Journal* 27, no. 3: 557–77.

Galindo-Rueda, F., and A. Vignoles. 2007. The heterogeneous effect of selection in UK secondary schools. In *Schools and the equal opportunity problem*, ed. L. Woessmann and P.E. Peterson, 103–28. Cambridge, MA: MIT Press.

Garces, E., D. Thomas, and J. Currie. 2002. Longer-term effects of head start. *American Economic Review* 92, no. 4: 999–1012.

Garibaldi, P., F. Giavazzi, A. Ichino, and E. Rettore. 2007. College cost and time to complete a degree: Evidence from tuition discontinuities. NBER Working Paper 12863. Cambridge, MA: National Bureau of Economic Research.

Gibbons, S., and S. Machin. 2006. Paying for primary schools: Admission constraints, school popularity or congestion? *Economic Journal* 116, no. 510: C77–92.

Gibbons, S., S. Machin, and O. Silva. 2008. Choice, competition, and pupil achievement. *Journal of the European Economic Association* 6, no. 4: 912–47.

Griliches, Z. 1979. Siblings models and data in economics: Beginnings of a survey. *Journal of Political Economy* 87, no. 5: 37–64.

Haegeland, T., O. Raaum, and K.G. Salvanes. 2008. Pennies from heaven? Using exogeneous tax variation to identify effects of school resources on pupil achievements. IZA Discussion Paper 3561. Bonn: Institute for the Study of Labor.

Hanushek, E.A. 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, no. 2: 143–63.

Hanushek, E.A., J.F. Kain, J.M. Markman, and S.G. Rivkin. 2003. Does peer ability affect student achievement? *Journal of Applied Econometrics* 18, no. 5: 527–44.

Hanushek, E.A., J.F. Kain, S.G. Rivkin, and G.F. Branch. 2007. Charter school quality and parental decision making with school choice. *Journal of Public Economics* 91, no. 5: 823–48.

Hanushek, E.A., and M.E. Raymond. 2004. The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association* 2, nos. 2–3: 406–15.

Hanushek, E.A., and L. Woessmann. 2006. Does early tracking affect educational inequality and performance? Differences-in-differences evidence across countries. *Economic Journal* 116, no. 510: C63–76.

Harmon, C., and I. Walker. 1995. Estimates of the economic return to schooling for the United Kingdom. *American Economic Review* 85, no. 5: 1278–86.

Heckman, J.J., S.H. Moon, R. Pinto, P.A. Savelyev, and A. Yavitz. 2010. The rate of return of the High/Scope Perry Preschool Program. *Journal of Public Economics* 94, nos. 1–2: 114–28.

Hoxby, C.M. 2000a. Does competition among public schools benefit students and taxpayers? *American Economic Review* 90, no. 5: 1209–38.

Hoxby, C.M. 2000b. The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics* 115, no. 4: 1239–85.

Hoxby, C.M. 2007. Does competition among public schools benefit students and taxpayers? Reply. *American Economic Review* 97, no. 5: 2038–55.

Ichino, A., and R. Winter-Ebmer. 2004. The long-run educational cost of World War II. *Journal of Labor Economics* 22, no. 1: 57–86.

Jacob, B.A. 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics* 89, nos. 5–6: 761–96.

Jürges, H., K. Schneider, and F. Büchel. 2005. The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association* 3, no. 5: 1134–55.

Kremer, M. 2003. Randomized evaluations of educational programs in developing countries: Some lessons. *American Economic Review* 93, no. 2: 102–6.

Krueger, A.B. 1999. Experimental estimates of education production functions. *Quarterly Journal of Economics* 114, no. 2: 497–532.

Krueger, A.B., and P. Zhu. 2004. Another look at the New York City school voucher experiment. *American Behavioral Scientist* 47, no. 5: 658–98.

Lavy, V. 2009. Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review* 99, no. 5: 1979–2011.

Lavy, V. Forthcoming. Mechanisms and effects of free choice among public schools. *Review of Economic Studies.*

Leuven, E., M. Lindahl, H. Oosterbeek, and D. Webbink. 2007. The effect of extra funding for disadvantaged students on achievement. *Review of Economics and Statistics* 89, no. 4: 721–36.

Leuven, E., H. Oosterbeek, and M. Rønning. 2008. Quasi-experimental estimates of the effect of class size on achievement in Norway. *Scandinavian Journal of Economics* 110, no. 4: 663–93.

Ludwig, J., and D.L. Miller. 2007. Does head start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics* 122, no. 1: 159–208.

Machin, S.J., and S. McNally. 2007. New technology in schools: Is there a payoff? *Economic Journal* 117, no. 522: 1145–67.

Machin, S.J., and S. McNally. 2008. The literacy hour. *Journal of Public Economics* 92, nos. 5–6: 1441–62.

Manning, A., and J.-S. Pischke. 2006. Comprehensive versus selective schooling in England and Wales: What do we know? IZA Discussion Paper 2072. Bonn: Institute for the Study of Labor.

Manski, C.F. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economics and Statistics* 60, no. 3: 531–42.

Meghir, C., and M. Palme. 2005. Educational reform, ability and family background. *American Economic Review* 95, no. 1: 414–24.

Messer, D., and S.C. Wolter. 2009. Money matters: Evidence from a large-scale randomized field experiment with vouchers for adult training. CESifo Working Paper 2548. Munich: CESifo.

Metzler, J., and L. Woessmann. 2009. The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. Mimeo, University of Munich.

Muralidharan, K., and V. Sundararaman. 2009. Teacher performance pay: Experimental evidence from India. NBER Working Paper 15323. Cambridge, MA: National Bureau of Economic Research.

Oreopoulos, P. 2006. Estimating average and local average treatment effects when compulsory schooling laws really matter. *American Economic Review* 96, no. 1: 152–75.

Pekkarinen, T., R. Uusitalo, and S. Kerr. 2009. School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics* 93, nos. 7–8: 965–73.

Peterson, P.E., and W.G. Howell. 2004. Efficiency, bias, and classification schemes: A response to Alan B. Krueger and Pei Zhu. *American Behavioral Scientist* 47, no. 5: 699–717.

Peterson, P.E., W.G. Howell, P.J. Wolf, and D.E. Campbell. 2003. School vouchers: Results from randomized experiments. In *The economics of school choice*, ed. C.M. Hoxby, 107–44. Chicago, IL: University of Chicago Press.

Puhani, P.A., and A.M. Weber. 2007. Does the early bird catch the worm? Instrumental variable estimates of early educational effects of age of school entry in Germany. *Empirical Economics* 32: 359–86.

Rivkin, S.G., E.A. Hanushek, and J.F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73, no. 2: 417–58.

Rockoff, J.E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94, no. 2: 247–52.

Rothstein, J. 2007. Does competition among public schools benefit students and taxpayers? Comment. *American Economic Review* 97, no. 5: 2026–37.

Sacerdote, B. 2001. Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics* 116, no. 2: 681–704.

Schneider, B., M. Carnoy, J. Kilpatrick, W.H. Schmidt, and R.J. Shavelson. 2007. *Estimating causal effects using experimental and observational designs: A think tank white paper.* Washington, DC: American Educational Research Association.

Schwerdt, G., and A.C. Wuppermann. 2009. Is traditional teaching really all that bad? A within-student between-subject approach. CESifo Working Paper 2634. Munich: CESifo.

Stock, J.H., and M.W. Watson. 2006. *Introduction to econometrics.* 2nd ed. Boston, MA: Addison-Wesley.

Todd, P.E., and K.I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *Economic Journal* 113, no. 485: F3–33.

Vigdor, J., and T. Nechyba. 2007. Peer effects in North Carolina public schools. In *Schools and the equal opportunity problem*, ed. L. Woessmann and P.E. Peterson, 73–101. Cambridge, MA: MIT Press.

Webbink, D. 2005. Causal effects in education. *Journal of Economic Surveys* 19, no. 4: 535–60.

West, M.R., and P.E. Peterson. 2006. The efficacy of choice threats within school accountability systems: Results from legislatively-induced experiments. *Economic Journal* 116, no. 510: C46–62.

West, M.R., and L. Woessmann. 2010. 'Every Catholic child in a Catholic school': Historical resistance to state schooling, contemporary school competition, and student achievement across countries. *Economic Journal* 120, no. 546: F229–55.

Woessmann, L. 2005. Educational production in Europe. *Economic Policy* 20, no. 43: 445–504.

Woessmann, L., and M.R. West. 2006. Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review* 50, no. 3: 695–736.

Zimmerman, D.J. 2003. Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics* 85, no. 1: 9–23.