6

Quantitative Research Methods in Education Finance and Policy

Patrick J. McEwan

INTRODUCTION

Researchers in education finance and policy rely on a myriad of quantitative methods. The most common include regression analysis, a mainstay of social science research, but they increasingly include experimental or quasi-experimental methods (McEwan & McEwan, 2003; Shadish, Cook, & Campbell, 2002). These methods are particularly suited to addressing research questions about the causal relationship between a policy or program and education outcomes.[1] Do school finance reforms increase the equity of school expenditures? Does attending a private instead of a public school improve students' achievement? Does financial aid increase the probability that students attend college? This chapter describes a range of quantitative methods that can be used to address causal questions, placing special emphasis on the methods' rationale, intuition, and pitfalls.

METHODS FOR WHAT?

Defining Research

In 2002, following a request of the U.S. Department of Education, the National Research Council (NRC) defined principles of scientific inquiry in education research (Shavelson and Towne, 2002, pp. 54–73). To paraphrase the NRC report, education research should: (1) pose important questions; (2) investigate them empirically, with appropriate research methods; (3) provide an explicit answer, after assessing and discarding plausible alternatives; (4) replicate and generalize the answers across contexts; and (5) do so in a transparent and professionally accountable way. The greatest controversy lies in the definition of "important" research questions, which shapes views on "appropriate" methods. Shavelson and Towne (2002) define three categories of questions: (1) descriptive (what is happening?), (2) causal (does it work?), and (3) process (how does it work?).

Descriptive research establishes or refutes patterns in the data, inspires theoretical explanations of the observed facts, guides the design of causal research, and provides better context for interpreting and generalizing causal results. Causal research tests for cause-and-effect relationships, rather than mere correlations, between policy interventions and policy outcomes. It helps

test, refine, and possibly discard theoretical explanations of empirical regularities in education. Most practically, it helps policy makers assess the relative merits of interventions and allocate resources to better ones, however defined. Process research inquires why a policy intervention does or does not affect policy outcomes. Did a single component of the intervention not work? Was it implemented well or poorly? Did teachers, students, or other participants respond by altering their behavior? To the extent that it unpacks the mechanisms explaining a "black box" causal effect, process research helps establish whether the same result can be replicated elsewhere.

Choosing Research Methods

The choice of a research method hinges on the question posed. Descriptive research employs standard quantitative techniques (Tukey, 1977) to describe the central tendency and dispersion of single variables (e.g. school expenditures or teacher characteristics) and their statistical association with other variables (e.g. student characteristics or test scores). Descriptive research also employs qualitative techniques—including case studies, ethnographies, or pragmatic reportage—to describe classrooms, schools, legislation, and political contexts.

Causal research uses mostly quantitative techniques, especially variants of regression analysis, to isolate the unique contribution of a policy to an education outcome. Regression methods are increasingly combined with, or supplanted by, experimental or quasi-experimental research methods, described below. Causal research employs qualitative methods less commonly, despite the potential applications (King, Keohane, & Verba, 1994).

Finally, process research is conducted with mixed, but frequently qualitative methods. In the best of cases, process research is built into an existing causal research designs. For example, researchers have employed qualitative methods, especially teacher interviews and classroom observations, to explain the large or small effects of reduced class sizes (Zahorik et al., 2000), Catholic school attendance (Bryk, Lee, & Holland, 1993), and whole-school reform (Cook et al., 1999; Cook et al., 2000).

The Emerging Importance of Causal Research

This chapter emphasizes quantitative methods for causal inference. In particular, it discusses regression analysis and related statistical techniques, as well as increasingly popular experimental and quasi-experimental research designs for analyzing the causal impact of policy interventions. It does so because these methods are foundational in modern research on education finance and policy, as evidenced by other chapters in the *Handbook*. Even so, some important methodological innovations have been slow to filter down to the day-to-day practice of education policy researchers, in part because the methods are confined to specialized or highly technical journal articles.

At least two factors have spurred the growing interest in valid causal research. First, policy makers are increasingly interested in policies or programs that have been rigorously shown to cause improvements in student outcomes. The interests of policy makers, in turn, have been shaped by an increasing federal emphasis on "scientific" research in education (Shavelson & Towne, 2002; U.S. Department of Education, 2003). For example, the U.S. Department of Education allocates grants to state and local education agencies for the implementation of reading programs, but only if they have been shown to "work."[2]

Second, education researchers have access to increasingly rich data sources that facilitate the application of additional methods (for examples, see Dee, Evans, & Murray, 1999; Loeb & Strunk, 2003). The data fall into three broad categories: (1) national samples of students;[3] (2) state and local administrative data that include repeated observations on a population of students;[4] and

(3) site-specific, fieldwork-based data such as the Tennessee STAR class size experiment. The second and third categories have been a particular catalyst to recent education policy research, because they facilitate better methods described below (Loeb & Strunk, 2003).

What This Chapter Does Not Address

Given the chapter's limited scope, it does not address three issues. First, the chapter does not discuss the application of quantitative methods to descriptive- or process-oriented research questions. Second, the chapter does not describe qualitative methods, despite their evident or potential merits in addressing descriptive, causal, or process questions. Third, it does not discuss a few quantitative methods that, while used by policy researchers, are the subject of their own specialized literatures. These include cost-benefit and cost-effectiveness analysis (Levin and McEwan, 2001; 2002), meta-analytic techniques for summarizing the results of many studies (Cooper & Hedges, 1994), and methods of measuring and comparing student achievement (Crocker & Algina, 1986; Kolen & Brennan, 2004).

DEFINITIONS OF RESEARCH TERMS

The results of cause-testing research, regardless of the method, are commonly judged by two criteria: *internal validity* and *external validity* (Cook & Campbell, 1979; Meyer, 1995; Shadish et al., 2002). A result is internally valid when it identifies a believable causal link between a policy or program and an education outcome. (The policy or program is often generically called a *treatment.*) A causal result is externally valid when it can be generalized to modified versions of the policy treatment, to alternate measures of student outcomes, to diverse populations of students or schools, or to different policy contexts. Much of the following discussion—and recent policy debates—emphasize the importance of improving internal validity (Barrow & Rouse, 2005; U.S. Department of Education, 2003). However, an internally valid result may be of little use to policy makers if it cannot be usefully generalized.

To take one example, California's large-scale and costly class size reduction was inspired by small-scale research in Tennessee that demonstrated positive—and internally valid—effects on student achievement (Mosteller, 1995; Krueger, 1999). Unlike the Tennessee research, California's large-scale implementation substantially increased demand for teachers. Statewide, deteriorating teacher quality appears to have offset the positive effects of class size reduction, resulting in a substantially modified version of the policy (Jepsen & Rivkin, 2002). The California experience highlights the trade-offs that often exist between internal and external validity.

The causal effect of a policy treatment is the difference between students' outcomes when treated, and the same students' outcomes when not treated. The latter is simply called the *counterfactual.* Short of procuring a time machine, it cannot be observed because treatments cannot be undone. Instead, research methods are employed to "create reasonable approximations to the physically impossible counterfactual" (Shadish et al., 2002, p. 5).[5] Researchers typically estimate counterfactual outcomes by identifying a separate group of untreated students, called a *control group* or *comparison group* (the former term is sometimes reserved for randomized experiments, but is used loosely). Members of the control or comparison group should be similar to their treated counterparts, in every respect but for exposure to the treatment.

In practice, the groups are often dissimilar, in ways that affect the outcomes of interest but have nothing to do with the treatment (i.e., students' families have different incomes, because higher-income families were more or less likely to choose the treatment). If this occurs, then the

mean outcomes of comparison students incorrectly estimate the policy counterfactual. Thus, a simple difference in the average outcomes of treatment and comparison students can yield a misleading estimate of the treatment's causal effect, and is said to suffer from *selection bias*. Selection bias is a pervasive challenge to internal validity, and the research methods discussed below are aimed at lessening it (on related challenges, see Shadish et al., 2002; Meyer, 1995).

Researchers use two broad methods, sometimes in combination, to ensure that members of treatment and control/comparison groups are similar, on average. First, they make statistical controls for observed differences between students, often using regression analysis. Second, they influence how students are assigned to treatment and control groups. Cause-testing research is often lumped into three broad categories: *experimental, quasi-experimental,* and *non-experimental.* The essential difference among categories is the degree of control exerted by the researcher over who is assigned to the policy treatment (whether students, teachers, schools, districts, or states), and who is assigned to the control/comparison group.

In experimental research—often referred to as a randomized, controlled trial (RCT)—the assignment is entirely determined by luck of the draw, as in a researcher's flip of a coin. In quasi-experimental research, the broadest category of research, assignment may contain elements of randomness or purposeful assignment by the researcher, but some might be due to the individual choices of students, parents, or administrators (called *selection*). In non-experimental research, the researcher exerts absolutely no influence, and assignment is entirely due to selection. When greater control is exerted, then causal results often possess greater internal validity. The following sections develop this point.

METHODS FOR ANSWERING CAUSAL QUESTIONS

Statistical Controls for Observed Variables

Suppose that researchers collect non-experimental data from students who attend either private or public schools, but were not encouraged or coerced to do so by researchers. Their causal question is whether attending private school improves test scores.[6] A naïve researcher would simply estimate the difference between the average test scores of private students and public students, and ascribe it to the causal effect of school type. But the difference could be explained by pre-existing differences in students that are the result of selection. For one, private students in tuition-paying schools have higher incomes, on average, which might be associated with higher test scores.

Regression Analysis. Regression analysis is the first line of defense against this kind of selection bias.[7] A basic regression model can be written as:

$$A_i = \beta_0 + \beta_1 P_i + \beta_2 X_i + \varepsilon_i,$$

where A represents the test score of each student in the entire sample (the subscript i might range from 1 to 1000); P indicates whether a student attends a private school ($P = 1$) or public school ($P = 0$);[8] X indicates the value of a control variable, like family income, that one wishes to hold constant; and ϵ is an error term unique to each student. The error term captures the notion that test scores vary, for unobserved reasons, even among students attending the same school type and with the same incomes.

Using the method of ordinary least squares, researchers estimate the regression coefficients β_0, β_1, and β_2. In the absence of controls for X, the estimate of β_1 would be interpreted as the average difference between private and public students' achievement (the naïve estimate from

before). Upon controlling for X, it is the average difference *holding constant family income*. The immediate question is whether the difference can now be interpreted as the causal effect of private school attendance.

The causal interpretation rests on an assumption of regression analysis: that private school attendance (P), controlling for X, is uncorrelated with positive or negative shocks in test scores captured in the error term, ϵ.[9] What could produce such correlations? Suppose the existence of an unmeasured variable, M, that gauges parent motivation. Further suppose that children of motivated parents disproportionately attend private schools (M and P are positively correlated, even controlling for X) and that the children of motivated parents obtain higher test scores (M and A are positively correlated, even controlling for X).

In regressions that do not control for M, the net result is that attending private schools tends to be accompanied by positive shocks in students' test scores, the (non-causal) influence of greater unobserved motivation among their parents. In this example, estimates of the coefficient β_1 would be "too big" because of selection bias, leading to overly optimistic causal conclusions about private school effects. Yet, selection bias could also work in the opposite direction, depending on the sign of partial correlations between the excluded variable(s), the dependent variables (A), and the key treatment variable (P).[10] Omitting variables like M creates no bias in estimates of β_1 if (1) the omitted variables are uncorrelated with A, or (2) the omitted variables are uncorrelated with P.[11]

In non-experimental research settings like this, researchers have few remaining options. One is to collect and control for additional variables in the regression, but that is cold comfort to users of existing data sets. Even controlling for hundreds of variables proves unconvincing in many contexts. For example, the archetypal non-experimental study in education policy regresses student test scores on student, family, and school variables.[12] The third category often includes a key policy variable such as class size or student/teacher ratio. Yet, test scores (say, at the end of grade 6) are the cumulative product of family and school "inputs" received by students from birth onwards (Todd & Wolpin, 2003). Data sets, even detailed longitudinal ones, can never fully and accurately measure all inputs. Variables like class size are probably correlated with unobserved variables that determine achievement, and estimates of their causal effects are unpredictably biased upward or downward.[13]

In test score regressions, researchers often resort to controlling for test score measurements taken at earlier moments in students' careers (say, at the beginning of grade 6). By controlling for pre-tests in so-called value-added regressions, researchers hope to implicitly control for all inputs that affected test scores until that moment, thereby reducing the scope of bias. Nonetheless, there is no guarantee that omitted variables during the sixth grade, or earlier ones not captured by error-ridden pre-tests, do not continue to bias estimates.[14]

Propensity Score Matching. An alternative and increasingly popular method of controlling for observed variables is propensity score matching (for recent examples in education policy, see Behrman, Cheng, & Todd, 2004; Black & Smith, 2004; Hong & Raudenbush, 2006; Shapiro & Trevino, 2004). Researchers first estimate a propensity score for each student (or other unit) in the sample (Rosenbaum & Rubin, 1983). The score is a predicted probability that students receive a treatment, given their observed characteristics. So, in the prior example, researchers would estimate probabilities, using a probit or logit regression, that students attend a private school, given their family income (X) and other observed variables thought to influence propensities.

Then, each private student is matched to a "similar" public student, based exclusively on values of their propensity scores.[15] If students cannot be matched to a counterpart, then they are discarded from the sample (this might happen, for example, if children of millionaires, or with

other observed characteristics, *always* attend private schools). Estimates of private school effects are based on comparisons of average outcomes across students in propensity-score matched treatment and control groups.

The method's virtues are at least twofold (Ravallion, 2005). First, it imposes no arbitrary assumption of linearity on the relationships between outcomes, policy variables, and other controls, as in the previous section's regression model. Second, it removes treated (or untreated) students from the sample that have no obvious "match" in the other group. Intuitively, the observed uniqueness of such students implies that they are also unique in unobserved ways that could introduce selection bias.

Yet, like regression analysis, the causal interpretation of propensity score matching results rests on the unverifiable assumption that no unobserved variables are correlated with outcomes and with the probability of receiving a treatment. In this regard, it is no panacea for causal research. Some empirical comparisons suggest that linear regression analysis and propensity score matching yield similar results (Godtland et al., 2004; Vandenberghe & Robin, 2004).[16]

Randomized Assignment

In the 1990s, researchers in education policy grew disenchanted with the ability of statistical controls for observed variables to eliminate selection bias in non-experimental data.[17] A turning point was the widespread analysis and debate of results from a randomized experiment in Tennessee that identified the causal effect of smaller class sizes on student test scores.[18] At the time, Krueger (1999) opined that "one well-designed experiment should trump a phalanx of poorly controlled, imprecise observational studies based on uncertain statistical specification" (p. 528). His opinion reflected a broader movement in empirical economics to focus less on the rote application of statistical tools, and more on the quality of counterfactual reasoning deployed by researchers (Angrist & Krueger, 1999; Glewwe & Kremer, 2006).[19]

In the classic instance of randomized assignment, researchers flip a coin to determine which students are treated, and which are not (note that teachers, schools, districts, or even entire towns could also be randomly assigned). Thus, each student's probability of receiving the treatment is an identical 0.5.[20] The virtue of this approach is that it balances, by design, the distribution of students' observed and unobserved characteristics across treatment and control groups. The two groups are not identical, of course, but they should be similar, on average.[21] Because control group members are similar, except for their exposure to a treatment, they provide an ideal counterfactual estimate of outcomes. For this reason, randomized experiments are commonly asserted to be the "gold standard" method of answering causal research questions.

To obtain an internally valid estimate of causal effects, one estimates the mean difference between the outcomes of treated and untreated units (a pleasant irony of experiments is that credible causal conclusions are obtained with unsophisticated statistical methods). One could further apply regression analysis to control for observed differences between groups. If randomization proceeded without a hitch, then doing so is not strictly necessary to eliminate selection bias,[22] although it reduces the standard errors of estimates of causal effects.

Besides class size, randomized experiments have been used to explore the effects of multiple policy treatments on student test scores. These include teacher performance incentives (Glewwe, Ilias, & Kremer, 2003), whole-school reform (Borman et al., 2005; Cook, et al., 2000), and private school vouchers in New York City (Howell & Peterson, 2002; Krueger & Zhu, 2004), among many other topics.[23]

In the New York voucher experiment, it was not feasible to randomly assign students to attend a private-school. Rather, researchers randomly awarded private school tuition vouchers

to some students, and the treatment consisted of a "voucher offer." In practice, not all students offered a voucher actually used it to attend a private school; by the same token, some students denied a voucher still attended a private school. This highlights a common feature of almost all social experiments: a subset of randomly assigned participants do not comply with the initial assignment.

This is not a fatal flaw of the research design. One option is simply to compare the outcomes (test scores, in this case) of the full treatment and control groups as initially assigned, regardless of whether or not they take up the voucher offer. This yields an unbiased estimate of the aptly-named "intent-to-treat." Though not an estimate of the effect of actually using a voucher to attend a private school, it provides valuable information to policy-makers. To further recover the effect of using a voucher—the effect of the "treatment-on-the-treated"—researchers can use additional methods, including instrumental variables methods discussed in a later section.

Randomized experiments are not without pitfalls (for varied opinions, see Burtless, 1995; Heckman, 1995). One common critique is that attrition from treatment or control groups could re-introduce selection bias into experimental estimates. In a voucher experiment, for example, one might be concerned that students leave the experimental sample, perhaps because they did not receive a voucher offer and chose to move to a district outside the research site. Attrition creates bias in experimental mean comparisons, to the extent that differential attrition across treatment and control groups changes the balance of observed and unobserved characteristics (i.e., all higher income students in the control group leave the sample to attend private schools in another city). To be fair, attrition is a widespread problem in social science research, and is not confined to randomized experiments.

A second critique is that experiments, especially small-scale ones, yield causal conclusions of limited external validity (Shadish et al., 2002). As one example, it is unlikely that treatments affect all students similarly. A typical randomized experiment identifies the average causal effect among heterogeneous students (some of whom are strongly affected, and others not at all). Researchers with large enough samples can estimate causal effects within subsamples of students, perhaps dividing them by location, income, or race. Yet, even average effects in experimental samples may not be generalizable to the average student in the entire population, since initial samples are not always a random draw. In fact, many experiments pragmatically begin with volunteers, whether students (Howell & Peterson, 2002) or schools (Cook et al., 2000). Results from volunteer students or schools may be harder to generalize to the broader population.

Though increasingly common, one might ask why "gold-standard" experiments are not used more often in education research (Cook, 2002). In education finance, some policy treatments are not amenable to randomized assignment. These include, for example, revised formulas for collecting and distributing education revenues that are imposed by courts or state legislatures. The close links of researchers to state-specific policy environments, and a desire to maximize the external validity of results, have focused attention on quasi-experimental methods for causal inference.

Other policies, such as test-based accountability, reward teachers and schools for their causal effect, or lack thereof, on student achievement. In such instances, randomized assignment can *never* be practically used to obtain such effects among the entire population of public schools, teachers, and students. As a consequence, recent state accountability laws rely on a mélange of quasi- and non-experimental approaches to assess teachers' and schools' "value-added."

Discontinuity Assignment

One of the most credible quasi-experimental methods is the regression-discontinuity design (RDD).[24] In the RDD, researchers assign students to treatment or control groups on the basis of

a single assignment variable—often a test score, but potentially any continuous variable—and a specified cutoff value. To provide an illustration, suppose that a thousand students vie for college financial aid by taking a pre-test (the assignment variable). Students with scores of 50 or above (the cutoff) receive aid, and those with scores below 50 do not. Note that assignment is not randomized, as in the flip of coin, but neither is it due to selection. This provides sufficient leverage to identify the causal effect of financial aid on some students' subsequent outcomes.

The causal effect is estimated by comparing the outcomes of treatment and control students whose values of the assignment variables are close to 50.[25] The intuition is that such students should be very similar, not just in their values of the pre-test, but in other observed and unobserved ways. At the very least, observed and unobserved characteristics of the students should not vary *sharply* in the vicinity of the cutoff. In short, control students (just to the left of the cutoff) provide a good counterfactual estimate of outcomes for treated students (just to the right). Thus, any sharp—or discontinuous—changes in outcomes near the cutoff can be attributed to the financial aid treatment.

The intuition of this approach is best understood with a visual analogy. In the absence of any treatment, suppose that one graphed a scatterplot of a college post-test on the y-axis against the pre-test on the x-axis. The scatterplot and a best-fitting line would likely indicate a positive relationship. They key point is that one would not anticipate a sharp break or discontinuity in the absence of the treatment. When the treatment is applied, perhaps accompanied by a break in outcomes near the cutoff, one's confidence is bolstered that it has a causal interpretation.

Among a growing litany of topics in education policy, the RDD has been applied to estimate the effects of class size reduction (Angrist & Lavy, 1999; Urquiola, 2006), college financial aid (Kane, 2003; van der Klaauw, 2002), early childhood education (Gormley & Gayer, 2005; Ludwig & Miller, 2007), teacher training (Jacob & Lefgren, 2004), and compensatory education for disadvantaged children (Chay et al., 2005).

A hallmark of recent papers is that researchers do not specify cutoffs or implement the assignment process. Instead, researchers take advantage of cutoff-based assignment that administrators used to allocate resources in a transparent, fair, or efficient way (i.e., needy or meritorious students receive financial aid, low-scoring schools receive assistance or sanctions and high-scoring ones receive rewards, less-effective teachers receive training, and so on). The unintended usefulness of such rules to researchers has only recently been noted in many cases, even when discontinuity assignment has a long history, as in the Head Start program (Ludwig & Miller, in press).

What are the potential pitfalls of using discontinuity-based assignment? The most serious, related to internal validity, is that students, or others subject to discontinuity assignment, are familiar with the potential intervention, the assignment variable, and the value of the cutoff. If they have incentives to receive the treatment, or not, then they may well attempt to manipulate their values of the assignment variable (Lee, in press; McCrary, in press). As in the non-experimental context, the concern is that manipulation may introduce selection bias into estimates of causal effects. For example, suppose that students with rich parents are aware of financial aid assignment rules and obtain extra pre-test tutoring for their children. The result is that treated children, just to the right of the cutoff, also happen to be somewhat wealthier, and perhaps more likely to attend college even without the treatment.

Precise manipulation of many continuous assignment variables is actually harder than it might seem (Lee, in press). While most families can probably influence their child's pre-test score, random errors in testing make it unlikely that they can affect it within a very narrow band of scores around the assignment cutoff. Students within this narrow band contribute the most to estimates of causal effects, implying that assignment variable manipulation must be very precise to bias regression-discontinuity effects.[26] To test for manipulation, researchers typically search

for suspicious clustering of students on either side the cutoff (McCrary, in press). They also compare students' observed characteristics near the cutoff, which should vary smoothly across the break, in the absence of manipulation.

Instrumental Variables

In the majority of cases, the assignment of students, schools, or other units to policy treatments is neither random nor based on values of an observed assignment variable. Besides controlling for observed characteristics, what remaining methods are available to identify the causal effect of policies on outcomes? One of most popular in the last decade has been instrumental variables (see Wooldridge, 2002, 2006; Angrist & Krueger, 1999).

In non-experimental data, the receipt of policy treatments is usually correlated with unobserved characteristics of individuals that affect outcomes. Therein lies the empirical dilemma. Yet, some individuals in the sample might receive a treatment because of luck or because they were encouraged to do so for reasons unrelated to outcomes. The challenge is to base estimates of causal effects entirely on "clean" variation in treatment status—that is, variation uncorrelated with unobserved characteristics that affect outcomes. It is easier said than done.

One must identify an instrumental variable, or instrument, that fulfills two conditions (Bound, Jaeger, & Baker, 1995; Wooldridge, 2002, 2006). First, it must be strongly correlated with the probability of receiving an intervention. This condition, straightforward to test in the data, is needed to ensure that the instrument actually induces students to alter their treatment status. Second, the instrument cannot be correlated with unexplained variation in the outcome variable (that is, the variation in outcomes that remains after controlling for other independent variables). The validity of the second assumption, more complicated to empirically test, usually rests upon the compelling reasoning of the researcher.

In applications to education policy, instruments are often related to features of geography or students' location, which are assumed to be "random" in some regard, and thus viable candidates to fulfill the second condition. Towards estimating private school effects, Figlio and Ludwig (2000) show that the availability of subway transportation in metropolitan areas affects the probability that families, especially poorer ones, choose private schools. Using this as an instrumental variable, their analysis suggests that private school attendance has strong effects on reducing some risky teenage behaviors. Their analysis must assume that transportation availability, of the instrument, is uncorrelated with student outcomes, controlling for other variables like family income.

Hoxby (2000) estimates the effects of competition among public school districts on students' outcomes like test scores.[27] She measures competition as the concentration of public school districts within metropolitan areas, where areas dominated by a few districts are assumed to be less competitive. The measure of competition is likely correlated with unobserved features of local students, schools, and communities that affect test scores. Hoxby argues that the number of streams in metropolitan areas (the instrumental variable) increases competition, because higher transportation costs led many areas to fragment into smaller school districts. The IV results suggest that metropolitan competition (induced by streams) has strong effects on test scores, based on the assumption that local geography is not correlated with unexplained test scores. Rothstein (2005) critiques the assumption that the instrumental variable is uncorrelated with unexplained student outcomes, as well as the measurement of the streams variable.

In each example, the validity of the second condition is hard to prove, and counter-examples easy to invent. (Do metropolitan areas with extensive subways have progressive mayors that invest in public schools? Do metropolitan areas with many streams and districts also have greater

segregation by race or socioeconomic status that lowers achievement?) In the most convincing IV analyses, there are a priori reasons to believe that instruments are uncorrelated with unexplained outcomes.

Difference-in-Differences

Difference-in-differences (DD) methods attempt to control for unobserved variables that bias estimates of causal effects, aided by longitudinal data collected from students, school, districts, or states. Researchers employ two varieties of longitudinal data. Panel data track the progress of the same students or teachers in successive months or years. Repeated cross-section data follow *different* groups of individuals (e.g., second-graders in successive years) that are clustered within the same schools, districts, or states.

The logic of DD causal inference is best communicated with an example based on repeated-cross section data (for its empirical implementation, see Dee & Levine, 2004). Of two states, Massachusetts and Maine, suppose that the former implements a finance reform—increasing state financing of local public school districts—and the latter does not. To estimate the reform's impact on district outcomes in Massachusetts, a naïve approach would compare outcomes across states, within a single year of post-reform data. The comparison is likely biased by selection, since unobserved differences across states could also affect outcomes.

Now consider the same comparison of outcomes, but within an earlier, pre-reform year of data. Evidently any differences in outcomes cannot be attributed to a Massachusetts reform that has yet to occur. Pre-reform differences in outcomes are perhaps due to unobserved differences across states that contaminated the previous, naïve estimate. To control for these unobserved variables, the DD estimate of the reform's effect subtracts the second difference from the first. The remaining "difference-in-differences" could be plausibly attributed to the reform. For this to be credible, the *change* in Maine's outcomes must be a good counterfactual for Massachusetts'. Yet, suppose that Massachusetts' outcomes rose more quickly than other states, even before the reform, because of economic growth due to a strong biotech industry. The DD will nonetheless attribute faster outcome growth in the treated state to the causal effect of reform.

In light of this pitfall, one of the best ways to assess the internal validity of DD results is to compare the trends of outcome variables across treatment and control groups *before* application of the treatment (Angrist & Krueger, 1999; Meyer, 1995). Evidence of similar trends bolsters confidence in the DD assumption. Dee and Levine (2004) estimated the effect of Massachusetts' state finance reform on districts' per-pupil state revenues. As controls, they used districts in Maine and Connecticut, which did not apply reforms, over the same pre- and post-reform period. DD estimates showed significant effects of the reform on local revenues. To support the use of these comparison groups, they showed that the outcome variables had similar trends in the three states in years prior to the reform.

There are many variants of DD analyses, depending on the context, research question, and data.[28] Dynarski (2003) estimates the effect of government-provided college subsidies on college attendance. She takes advantage of the fact that Social Security Administration used to provide large college subsidies to children with deceased parents, but abruptly stopped doing so, beginning with the high school class of 1982. Dynarski identifies students with deceased parents—the treatment group—who graduated from high school before and after the change, and students without deceased parents—a comparison group never eligible for the benefits—over the same period. The DD estimates suggest large effects of subsidies on college attendance.

Researchers increasingly apply DD methods to student-level panel data on test scores, applying a similar logic of causal inference. Some students' outcomes are observed before and after

exposure to a treatment. Their outcomes are compared to students never exposed to the treatment. Rouse (1998) provides a lucid example of this approach in her re-analysis of data from the Milwaukee voucher program. The treatment group consists of students, observed before and after their selection to receive a private school voucher. The comparison groups consist of (1) students who were denied a private school voucher or (2) students in public schools who never applied for one. The DD estimates suggest that treated students have faster gains in math scores, but not in reading, than students in both comparison groups. The necessary assumption, as in previous analyses, is that treated students would have had trends in achievement similar to untreated students in the absence of the treatment.[29]

Combining Methods to Improve Causal Inference

Researchers often apply multiple methods in the same study to bolster confidence in causal results. Almost every study employs statistical controls for family and student characteristics that affect outcomes. In experiments, the RDD, and a few DD applications, controls are not essential since careful control group selection alone should be enough to remove selection bias.[30] However, the further inclusion of controls in such studies provides a handy check of internal validity, since it should not substantially alter estimates of causal effects.[31]

Researchers often combine DD methods with experiments (Krueger & Zhu, 2004; Skoufias, 2005), the RDD (Chay et al., 2005; Jacob & Lefgren, 2004), and IV (Kuziemko, 2006; Loeb & Page, 2000). In experiments and the RDD, there should be no initial difference in pre-treatment outcomes across treatment and control groups, so using longitudinal data is not strictly necessary to control for selection bias.[32] But again, it provides a useful check of internal validity, and might improve the internal validity of research with a great deal of sample attrition.

Finally, researchers combine IV methods with randomized experiments and the RDD, especially to address imperfect compliance of students with random or cutoff-based assignment to policy treatments.[33] Returning to the previous example of New York's voucher experiment, students were randomly assigned to receive a voucher offer, but not all students accepted the offer and actually attended a private school. To recover an estimate of the treatment-on-the-treated (i.e., the effect of actually attending a private school), researchers used the voucher offer as an instrument for private school attendance (Howell & Peterson, 2002; Krueger & Zhu, 2004). The instrument plausibly fulfills both conditions: (1) it is correlated with private school attendance, and (2) the initial random assignment of the offer ensures that it is not correlated with unexplained outcomes. The resulting IV estimate provides a credible estimate of the effect of private school attendance on those induced to accept it by the voucher offer.[34]

CONCLUSIONS

This chapter has described quantitative research methods used to estimate the causal effect of policies on education outcomes. Some, like regression analysis with non-experimental data, are ubiquitous but not always capable of delivering strong causal conclusions. Others, like experimental and discontinuity research designs, are increasingly common in education finance and policy, but are still relatively under-utilized.

Good causal research is a necessary but not sufficient condition for designing and implementing good policy, a point not always clear in recent debates (U.S. Department of Education, 2003). Notwithstanding this chapter's emphasis on rigorous causal research methods, it does not address methods for answering descriptive or process questions nor does it review qualitative

research methods (King, Keohane, & Verba,1994). Both can be eminently "scientific" (Shavelson & Towne, 2002) and deserve serious attention from newer generations of researchers in education finance and policy.

NOTES

1. For methodological reviews in education policy, see Angrist (2004), Barrow and Rouse (2005), Glewwe and Kremer (in press), Hanushek (2002), Ludwig (2001), and McEwan and McEwan (2003). Angrist and Krueger (1999), Meyer (1995), and Ravallion (2001; 2005) provide reviews in the context of social policy. Shadish, Cook, and Campbell (2002) and its predecessors (Campbell & Stanley, 1963; Cook & Campbell, 1979) are canonical references in evaluation research.
2. The federally funded "What Works Clearinghouse" sifts through education research and harshly judges its ability to derive valid causal inferences about the impact of education programs. Within the Department of Education, the grant-making Institute for Education Sciences favors research proposals that use research methods able to credibly demonstrate causal impacts, notably randomized experiments and regression-discontinuity designs.
3. National samples, notably NELS:88, have been collected by the National Center for Education Statistics (NCES) and are available from their Web site (http://www.nces.ed.gov).
4. These include data from Chicago Public Schools (Jacob & Lefgren, 2004), Texas (Hanushek et al., 2007), North Carolina (Bifulco & Ladd, 2006), Florida (Sass, 2006), and teachers in New York state (Lankford, Loeb, & Wyckoff, 2002).
5. Shadish et al. (2002) describe the history of the counterfactual reasoning. It has been formalized by statisticians, especially Donald Rubin (Holland, 1986), in a framework that has been adopted by econometricians (Wooldridge, 2002, pp. 603–607; Angrist, 2004; Ravallion, 2005), and applied in recent research on education finance and policy.
6. For reviews of similar research, see McEwan (2000), Ladd (2002), and Neal (2002).
7. In this chapter, regression analysis implies ordinary least-squares regression (OLS). For basic discussions, see Wooldridge (2006) or Stock and Watson (2003). In education policy, it is common to apply multilevel or hierarchical models (Raudenbush & Bryk, 2002; Somers, McEwan, & Willms, 2004) that model error components and account for the potential correlation of errors within classrooms, schools, communities, or states. In so doing, they avoid understating standard errors of coefficients and over-stating their statistical significance. (The models do not necessarily, as is sometimes assumed, remove selection bias.) Economists are more likely to report OLS coefficient estimates accompanied by adjusted Huber-White standard errors that allow for arbitrary correlations among units within clusters (Wooldridge, 2002). In comparisons, OLS with adjusted standard errors and other multilevel models yield similar results, though OLS with standard errors *not* adjusted for clustering can dramatically underestimate standard errors (Angeles & Mroz, 2001). This issue is not discussed further, but the research cited in this chapter generally reports cluster-adjusted standard errors.
8. This assumes a binary policy intervention (treated or not), though the discussion can be generalized to continuously-measured policy interventions (e.g. class sizes of 1 to 50).
9. Formally, the assumption is that $\text{cov}(P_i, \varepsilon_i) = 0$.
10. It is common for researchers to conjecture about the direction of bias, in concert with the oft-reasonable presumption that selection on unobserved characteristics, like motivation, might work in the same direction as selection on observed characteristics, like family income (e.g. Somers et al., 2004). For a rigorous application of this reasoning to the effects of private school attendance on test scores, see Altonji, Elder, and Taber (2005).
11. The goal of randomized experiments is to ensure that the second condition holds by design.
12. For reviews of such studies in the United States, see Greenwald, Hedges, and Laine (1996), Hanushek (1997; 2002), and Krueger (2003). Reviews of international studies include Fuller and Clarke (1994), Hanushek (1995), and Glewwe and Kremer (2006).
13. Urquiola (2006) documents that disadvantaged, rural students in Bolivia, by virtue of their location, are

more likely to attend smaller classes. Presuming that some features of "disadvantage" are unobserved, and lead to lower test scores, regression-based estimates of the causal effect of small-class treatments are biased towards finding no effect. For related arguments focusing on the United States, see Boozer and Rouse (2001).

14. See Hanushek (1986) and Barrow and Rouse (2005). Boardman and Murnane (1979) and Todd and Wolpin (2003) formally analyze conditions under which pre-test controls may eliminate bias in test score regressions.
15. One could literally match units "by hand," based on a small number of observed characteristics. Doing so becomes increasingly difficult as the number of observed characteristics and matching categories increases. Propensity score matching provides a solution to this "curse of dimensionality." Matching algorithms and related analyses are still a topic of debate; for reviews, see Ravallion (2005) and the citations therein.
16. As a counter-example Black and Smith (2004) find that regression and propensity estimates of the effects of college quality on wages are similar in a sample of men, but not women.
17. A growing literature finds that non-experimental statistical approaches, including regression and propensity score matching, do a poor job of replicating experimental results (Agodini & Dynarski, 2004; Glewwe et al., 2004; Glazerman, Levy, & Myers, 2003).
18. See Mosteller (1995) and Krueger (1999). In developing-country research, a similar role was played by the large-scale experimental evaluation of PROGRESA, a Mexican program that awarded cash payments to families in exchange for participating in health and education programs (Skoufias, 2005).
19. The opinion is not limited to researchers in education policy or economics. The statistician David Freedman (1991) remarked that "regression models make it all too easy to substitute technique for work" (p. 300), and called on researchers to expend more "shoe leather" in the pursuit of the convincing counterfactual reasoning and data.
20. A coin flip or similar mechanism is only the simplest approach to designing randomized assignment. The essential point is that students or other units have well-defined probabilities of being assigned to the treatment. On the design, implementation, and analysis of randomized experiments, see Orr (1999) and Duflo, Glennerster, and Kremer (2006).
21. Indeed, one of the key tests for determining whether randomization "worked" is to test, and hopefully not reject, the null hypothesis of no average differences in observed characteristics across treatment and control group units. If there are differences, it could indicate random, but unlikely noise (like the person who flips a coin and gets heads 10 times in row). More perniciously, it could indicate attempts to manipulate random assignment, which is most likely in settings where researchers do not administer random assignment, and where individuals or other units have incentives to be treated or untreated (McEwan & Olsen, 2007).
22. In terms of the previous regression framework, randomized assignment ensures that the treatment (e.g., P) is uncorrelated with the error term, ϵ.
23. For overviews of experimentation in the United States, see Borman (2002) and, in developing countries, Glewwe and Kremer (2006).
24. See Hahn, Todd, and van der Klauuw (2001), Lee (in press), and Shadish et al. (2002). On the design's use and independent discovery in several disciplines, see Cook (in press). Cook and Wong (in press) and Buddelmeyer and Skoufias (2003) compare experimental and RDD estimates of the same programs. They find close correspondence, suggesting that the internal validity of RDD results is high.
25. See Chay, McEwan, and Urquiola (2005) and van der Klaauw (2002). The sample of students (or other assigned units) is often small near cutoffs, limiting the statistical precision of comparisons. Thus, researchers frequently apply parametric or non-parametric regression analysis to larger samples of data, further away from the cutoff (on techniques, see van der Klaauw, 2002; McCrary & Royer, 2006). Regressions control for an indicator of eligibility (i.e., pre-test above the cutoff) and additional controls for smooth functions of the assignment variable, often a quadratic or cubic function (Chay et al., 2005).
26. Such manipulation is perhaps more likely when using discrete assignment variables that individuals or organizations—with incentives to obtain or avoid treatment—can precisely set. For example, many

researchers note that day of birth is an assignment variable, and that children born after a specified assignment cutoff date are subject to treatments, like delaying school enrollment by one year (McCrary & Royer, 2006; McEwan & Shapiro, in press). Of course, motivated parents can manipulate day of birth by cesarean section or induced labor. While such manipulation might seem unlikely, Dickert-Conlin and Chandra (1999) show that U. S. parents precisely time year-end births to obtain tax benefits.

27. Belfield and Levin (2002) summarizes the empirical literature on competition.
28. One of the most common uses in economics, and increasingly so in education finance and policy, is to use repeated cross-section data on all 50 states, during periods in which some states were exposed to reform (perhaps at different times), and other states were not. The strategy has been applied, for example, to the effects of state accountability reforms (Hanushek & Raymond, 2005) and state finance reforms (Murray, Evans, & Schwab, 1998).
29. The assumption is not always tenable in student panel data on test scores. Suppose that students are more likely to apply to charter schools after experiencing slower test score growth in public schools than other students. If they switch to charter schools, a DD estimate of charter school effects could mistake this pre-existing trend for part of a charter school "effect." With more than two years of data, researchers can implement more complicated models that compare changes in test score *gains* rather than levels (Bifulco & Ladd, 2006; Hanushek et al., 2007; Sass, 2006). Repeated cross-section studies have employed a similar strategy, controlling for unit-specific linear time trends (see, e.g., the state-level studies of Loeb & Page, 2000 and Hanushek & Raymond, 2005).
30. This is not the case in common applications of IV, since fulfilling the second condition (instruments uncorrelated with *unexplained* outcomes) is more likely to be fulfilled when researchers have used controls to already explain much of the variance in outcomes.
31. For applications to experiments, regression-discontinuity, and DD, see Krueger (1999), Chay et al., (2005), and Dynarski (2003), respectively.
32. In experiments, this includes the full treatment and control groups in experiments; in the RDD, it includes smaller groups close to the cutoff.
33. For examples of experiments, see Krueger and Zhu (2004) and Rouse (1998). On the RDD, see van der Klaauw (2002).
34. In the analogous RDD case, units do not always comply with their initial cutoff-based assignment to treatment or control groups. RDDs with perfect compliance are called "sharp" designs, and those with imperfect compliance are "fuzzy" (Shadish et al., 2002). In the earlier example, families whose child is ineligible for financial aid (by virtue of obtain a pre-test score below 50) may nonetheless lobby to have the decision overturned. Other families may turn down the assigned treatment. Researchers can apply IV (as described in van der Klaauw, 2002 or Chay et al., 2005) to estimate a local version of the effect of the treatment-on-the-treated.

REFERENCES

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention program. *Review of Economics and Statistics, 86*(1), 180–194.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy, 113,* 151–184.

Angeles, G., & Mroz, T. A. (2001). *A guide to using multilevel models for the evaluation of program impacts.* Measure Evaluation Working Paper 01–33. Chapel Hill, NC: Carolina Population Center, University of North Carolina.

Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy, 20*(2), 198–212.

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 3A). Amsterdam: Elsevier.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics, 114*(2), 533–575.

Barrow, L., & Rouse, C. E. (2005). *Causality, causality, causality: The view of education inputs and outputs from economics.* Working Paper 2005–2015. Chicago: Federal Reserve Bank of Chicago.

Behrman, J. R., Cheng, Y., & Todd, P. E. (2004). Evaluating preschool programs when length of exposure to the program varies: A nonparametric approach. *Review of Economics and Statistics, 86*(1), 108–132.

Belfield, C., & Levin, H. M. (2002). The effects of competition between schools on educational outcomes: A review for the United States. *Review of Educational Research, 72*(2), 279–341.

Bifulco, R., & Ladd, H. F. (2006). The impacts of charter schools on student achievement: Evidence from North Carolina. *Education Finance and Policy, 1*(1), 50–90.

Black, D. A., & Smith, J. A. (2004). How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics, 121*(1-2), 99–124.

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education, 52*(2), 113–121.

Boozer, M., & Rouse, C. (2001). Intraschool variation in class size: Patterns and implications. *Journal of Urban Economics, 50*(1), 163–189.

Borman, G. D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education, 77*(4), 7–27.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2005). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal, 42*, 673–696.

Bound, J., Jaeger, D. A., & Baker, R. (1995). Problems with instrumental variable estimation when the correlation between the instruments and endogenous explanatory variable is weak. *Journal of the American Statistical Association, 90*(430), 443–450.

Bryk, A. S., Lee, V. E., & Holland, P. B. (1993). *Catholic schools and the common good.* Cambridge, MA: Harvard University Press.

Buddelmeyer, H., & Skoufias, E. (2003). *An evaluation of the performance of regression discontinuity design on PROGRESA.* Discussion Paper No. 827. IZA.

Burtless, G. (1995). The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives, 9*(2), 63–84.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review, 95*(4), 1237–1258.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199.

Cook, T. D. (in press). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics, and economics. *Journal of Econometrics.*

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cook, T. D., Habib, F. N., Phillips, M., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999). Comer's School Development Program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36*(3), 543–597.

Cook, T. D., Murphy, R. F., & Hunt, H. D. (2000). Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal, 37*(2), 535–597.

Cook, T. D., & Wong, V. C. (in press). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique.*

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth: Harcourt Brace Jovanovich.

Dee, T. S., Evans, W. N., & Murray, S. E. (1999). Data watch: Research data in the economics of education. *Journal of Economic Perspectives, 13*(3), 205–216.

Dee, T. S., & Levine, J. (2004). The fate of new funding: Evidence from Massachusetts' education finance reforms. *Educational Evaluation and Policy Analysis, 26*(3), 199–215.

Dickert-Conlin, S., & Chandra, A. (1999). Taxes and the timing of births. *Journal of Political Economy, 107*(1), 161–177.

Dynarski, S. M. (2003). Does aid matter? Measuring the effect of student aid on college attendance and completion. *American Economic Review, 93*(1), 279–288.

Duflo, E., Glennerster, R., & Kremer, M. (2006). *Using randomization in development economics research: A toolkit.* Unpublished manuscript, MIT.

Figlio, D. N., & Ludwig, J. (2000). *Sex, drugs, and Catholic schools: Private schooling and non-market adolescent behaviors.* Working Paper No. 7990. Cambridge, MA: National Bureau of Economic Research.

Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological Methodology, 21,* 291–313.

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. *Review of Educational Research, 64*(1), 119–157.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science, 589*(1), 63–93.

Glewwe, P., & Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 2, pp. 945–1017). Amsterdam: Elsevier.

Glewwe, P., Ilias, N., & Kremer, M. (2003). *Teacher incentives.* Working Paper No. 9671. Cambridge, MA: National Bureau of Economic Research.

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: The case of flip charts in Kenya. *Journal of Development Economics, 74,* 251–268.

Godtland, E. M., Sadoulet, E., de Janvry, A., Murgai, R., & Ortiz, O. (2004). The impact of farmer field schools on knowledge and productivity: A study of potato farmers in the Peruvian Andes. *Economic Development and Cultural Change, 53,* 63–92.

Gormley, W. T., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's Pre-K program. *Journal of Human Resources, 40*(3), 533–558.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361–396.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201–209.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*(3), 1141–1177.

Hanushek, E. A. (1995). Interpreting recent research on schooling in developing countries. *World Bank Research Observer 10*(2), 227–246.

Hanushek, E. A. (1997). Assessing the effects of school resources of student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141–164.

Hanushek, E. A. (2002). Publicly provided education. In A. J. Auerbach & M. Feldstein (Eds.), *Handbook of Public Economics* (Vol. 4, pp. 2045–2141). Amsterdam: Elsevier.

Hanushek, E. A., Kain, J. F., Rivkin, S. G., & Branch, G. F. (2007). Charter school quality and parental decision making with school choice. *Journal of Public Economics, 91*(5–6), 823–848.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297–327.

Heckman, J. J. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives, 9*(2), 85–110.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Hong, G., & Raudenbush, S. W. (2006). Evaluating Kindergarten retention policy: A case study of causal inference for multilevel observation data. *Journal of the American Statistical Association, 101*(475), 901–910.

Howell, W. G., & Peterson, P. E. (2002). *The education gap: Vouchers and urban schools.* Washington, DC: The Brookings Institution Press.

Hoxby, C. M. (2000). Does competition among public schools benefit students and taxpayers? *American Economic Review, 90*(5), 1209–1238.

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources, 39*(1), 50–79.

Jepsen, C., & Rivkin, S. (2002). *What is the tradeoff between smaller classes and teacher quality?* Working Paper No. 9205. Cambridge, MA: National Bureau of Economic Research.

Kane, T. J. (2003). *A quasi-experimental estimate of the impact of financial aid on college-going.* Working Paper No. 9703. Cambridge, MA: National Bureau of Economic Research.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research.* Princeton, NJ: Princeton University Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114*(2), 497–532.

Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal, 113*(485), F34–F63.

Krueger, A. B., & Zhu, P. (2004). Another look at the New York City voucher experiment. *American Behavioral Scientist, 47*(5), 658–698.

Kuziemko, I. (2006). Using shocks to school enrollment to estimate the effect of school size on student achievement. *Economics of Education Review, 25*(1), 63–75.

Ladd, H. F. (2002). School vouchers: A critical view. *Journal of Economic Perspectives, 16*(4), 3–24.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37–62.

Lee, D. S. (in press). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics.*

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Levin, H. M., & McEwan, P. J. (Eds.). (2002). *Cost-effectiveness and educational policy.* Larchmont, NY: Eye on Education.

Loeb, S., & Page, M. E. (2000). Examining the link between teacher wages and student outcomes: The importance of alternative labor market opportunities and non-pecuniary variation. *Review of Economics and Statistics, 82*(3), 393–408.

Loeb, S., & Strunk, K. (2003). The contribution of administrative and experimental data to education policy research. *National Tax Journal, 56*(2), 415–438.

Ludwig, J. (2001). Problems in the estimation of school effects: Insights from improved models. In D. H. Monk, H. J. Walberg, & M. C. Wang (Eds.), *Improving educational productivity* (Vol. 1, pp. 209–236). Greenwich, CT: Information Age Publishing.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics, 122*(1), 159–208.

McCrary, J. (in press). Manipulation of the running variable in the regression-discontinuity design: A density test. *Journal of Econometrics.*

McCrary, J., & Royer, H. (2006). *The effect of maternal education on fertility and infant health: Evidence from school entry laws using exact date of birth.* Unpublished manuscript, University of Michigan.

McEwan, E. K., & McEwan, P. J. (2003). *Making sense of research.* Thousand Oaks, CA: Corwin.

McEwan, P. J. (2000). The potential impact of large-scale voucher programs. *Review of Educational Research, 70*(2), 103–149.

McEwan, P. J., & Olsen, R. (2007). *Admission lotteries in charter schools.* Unpublished manuscript, Wellesley College and Urban Institute.

McEwan, P. J., & Shapiro, J. (in press). The benefits of delayed primary school enrollment: Discontinuity estimates using exact birth dates. *Journal of Human Resources.*

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics, 13*(2), 151–161.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*(2), 113–127.

Murray, S. E., Evans, W. N., & Schwab, R. M. (1998). Education-finance reform and the distribution of education resources. *American Economic Review, 88*(4), 789–812.

Neal, D. (2002). How vouchers could change the market for education. *Journal of Economic Perspectives, 16*(4), 25–44.

Orr, L. L. (1999). *Social experiments.* Thousand Oaks, CA: Sage.

Ravallion, M. (2001). The mystery of the vanishing benefits: An introduction to impact evaluation. *World Bank Economic Review, 15*(1), 115–140.

Ravallion, M. (2005). *Evaluating anti-poverty programs.* Policy Research Working Paper No. 3625. Washington, DC: World Bank.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rothstein, J. (2005). *Does competition among public schools benefit students and taxpayers? A comment on Hoxby (2000).* Working Paper No. 11215. Cambridge, MA: National Bureau of Economic Research.

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics, 113*(2), 553–602.

Sass, T. R. (2006). Charter schools and student achievement in Florida. *Education Finance and Policy, 1*(1), 91–122.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shapiro, J., & Trevino, J. M. (2004). *Compensatory education for disadvantaged Mexican students: An impact evaluation using propensity score matching.* Policy Research Working Paper 3334. Washington, DC: World Bank.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Skoufias, E. (2005). *PROGRESA and its impacts on the welfare of rural households in Mexico.* Research Report No. 139. Washington, DC: International Food Policy Research Institute.

Somers, M.-A., McEwan, P. J., & Willms, J. D. (2004). How effective are private schools in Latin America? *Comparative Education Review, 48*, 48–69.

Stock, J. H., & Watson, M. W. (2003). *Introduction to econometrics.* Boston: Addison-Wesley.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113*(485), F3–F33.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural Bolivia. *Review of Economics and Statistics, 88*(1), 171–177.

U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide.* Washington, DC: U.S. Department of Education, Institute of Education Sciences.

van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review, 43*(4), 1249–1286.

Vandenberghe, V., & Robin, S. (2004). Evaluating the effectiveness of private education across countries: A comparison of methods. *Labour Economics, 11*, 487–506.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data.* Cambridge, MA: MIT Press.

Wooldridge, J. M. (2006). *Introductory econometrics: A modern approach* (3rd ed.). Thomson South-Western.

Zahorik, J., Molnar, A., Ehrle, K., & Halbach, A. (2000). Smaller classes, better teaching? Effective teaching in reduced-size classes. In S. W. M. Laine & J. G. Ward (Eds.), *Using what we know: A review of the research on implementing class-size reduction initiatives for state and local policymakers* (pp. 53–73). Oak Brook, IL: North Central Regional Educational Laboratory.