The Contribution of Administrative and Experimental Data to Education Policy Research

Susanna Loeb & Katharine Strunk
School of Education, Stanford University, Stanford, CA 94305-3084

National Tax Journal
Vol. LVI, No. 2
June 2003

INTRODUCTION

The last decade has seen a surge in empirical research by economists addressing the impact of school reform policies. This wave of research dates back to Card and Krueger (1992) and a series of analyses that look across the United States to test the effects of school inputs on outcomes such as achievement, educational attainment, and earnings. Many of these studies use either Census data (Heckman, Layne–Farrar, and Todd, 1996) or national longitudinal surveys (Betts 1996; Loeb and Bound, 1996; Grogger, 1996) to estimate input effects. Since then, much of the research has turned from the nation to states and cities, using local administrative and experimental data to focus on specific policy initiatives.[1] The shift in emphasis from the national to the local has enabled researchers to incorporate greater institutional and policy detail into their analyses. Local data often follow individuals over time, allowing researchers to use empirical techniques not possible with pooled cross–sectional Census data. At the same time, the administrative data provide deep coverage of local areas, which national surveys that sample only a small portion of any locality, do not. While local administrative data provide advantages over census and survey data for estimating the causal effects of some policies, experimental data alleviate the perennial concern that correlational analyses have not established causality. However, experiments are not easy to implement in education, they can't be used to address all types of education policy, and they tend to be too small to elicit the responses generated by large–scale reforms.

As data have improved and studies have proliferated, the recent literature is more characterized by seemingly inconsistent results than by increased consensus over the impact of education policy reforms. Estimates of the impact of similar policies across analyses, datasets, and localities often vary. For example, there appear to be inconclusive results surrounding the impact of class size on student achievement. Hoxby (2000a) finds no effect, while Angrist and Lavy (1999), who use similar techniques but different data, find positive

[1] Some earlier work, such as Ferguson (1991), also uses state and city level data.

effects of smaller classes, as do Krueger (1997) and Hanushek (1999) in their studies of the Tennessee STAR experiment. Some of the variation in the estimates may be due to inadequate statistical techniques, as each type of data and empirical approach has its own weaknesses, and no approach has been able to overcome all concerns. Yet much of the variation may not stem from econometric error but, instead from heterogeneity of policies, from the influence of local conditions on policy effects, and/or from the differential effects of a single policy on heterogeneous individuals. The answer to whether a policy improves student learning may not be a simple "yes" or "no." Instead, the answer is likely to depend on the context in which the policy is implemented, the scale of the reform, and the characteristics of the students in question. As a result, it is worth considering the findings individually and working to create a structure that allows for differences in context and that can be used to understand complex effects of policy reforms.

This paper reviews recent innovations in data usage in empirical studies of education policy effects, concentrating on large administrative data collection efforts and experiments. It then summarizes a number of the findings in three areas of education policy that have been the focus of recent research and debate: school finance, accountability, and choice. Although there have been fundamental reforms in education policy at both the preschool and higher education levels, this paper focuses on elementary and secondary education policy, paying special attention to the types of data used and the impact of the empirical approach on the findings.[2]

DATA

Recent research in education finance is characterized by the use of a wide range of data types including national longitudinal surveys, Census data, state and city administrative datasets, and experimental data. Each type of data has clear advantages and disadvantages. For example, while Census data provide national coverage and a large sample size, they lack the detail that can be obtained through smaller national surveys. Census data do not follow individuals over time, though they can be used to follow jurisdictions such as school districts, counties or states. State and city administrative data often do provide detail on schools and individuals and can trace them over time, but only within that city or state, making it difficult to extrapolate the findings to other locations. Only experiments are designed specifically to determine causality, however, they are difficult to implement effectively and are often limited in scope. Below we outline briefly the different sources of data and further advantages and disadvantages of each.

National Surveys and Census Data

Until recently, surveys provided the only data that included achievement outcomes and linked students to schools and teachers. Surveys still provide the best source of information on students over long periods of time. The National Longitudinal Survey of Youth (NLSY) (http://www.bls.gov/nls/home.htm), High School and Beyond (HS&B) (http://nces.ed.gov/surveys/hsb), and the National Education Longitudinal Study of 1988 (NELS88) (http://nces.ed.gov/sur-

[2] Since the 1996 Welfare reform, the federal government has increased subsidies for child care, largely in the form of vouchers to families. One federal study reports that over one million additional families now receive public child care support each year, compared to before the 1996 reforms (Collins et al., 2000). In higher education, many states have altered their finance system, some shifting from low–tuition policies to higher–tuition higher–aid policies, and some instituting incentive programs, such as the Georgia Hope Program, that tie aid to high school achievement.

veys/nels88) are examples of datasets that follow students through school and into the workforce. Stinebrickner (2002) uses the NLSY to assess who goes into teaching and why they leave, finding that most female teachers who stop teaching leave the workforce altogether, many to care for a newborn child. As is clear from Stinebrickner's work, survey datasets track individuals over long periods of time. They are also the best source for assessing a range of student outcomes such as homework completion, sports team participation, the quality and characteristic of the undergraduate institution selected, and volunteerism. Figlio and Ludwig (2000) use NELS88 to compare the sexual behavior, drug use, and gang activity of Catholic and public schools students, finding that Catholic schools reduce teen sexual activity, arrests, and hard drug use, but not drinking, smoking, gang involvement or marijuana use. Betts and Shkolnik (1999) use the Longitudinal Study of American Youth to look at the impact of class size on teaching behaviors, including group and individual instruction, time spent on discipline, and textbook coverage. They find that smaller classes lead to more individual instruction and less discipline, although the differences are minor. Rose and Betts (2001), using HS&B, find large effects of math curriculum on students' later outcomes. Those students who take higher–level math in high school are more likely to complete college and have higher incomes even after controlling for math ability.

Because these surveys follow students over time, researchers can use value–added models of student achievement. By looking at differences in the achievement gains of students across policy environments, instead of differences in achievement levels, value–added models reduce the bias caused by the sorting of higher or lower ability students. While this is preferable to data that have only levels of student achievement, it is still difficult to establish causality. Students with faster or slower learning trajectories may be sorted into specific policy regimes (such as larger classes). Because each survey tracks only a single cohort, researchers must use cross–sectional policy data in order to assess effects. Studies compare a policy (for example, spending per pupil) in one school to the policy in another school, making it difficult to separate school effects from policy effects. Some researchers have tried to link surveys in order to allow for some fixed effects for jurisdictions or institutions, but have run into difficulties because the surveys cover different schools and ask different questions (see Brewer, Eide, and Ehrenberg, 1999).

Unlike national survey data, Census data do allow for fixed effects at the jurisdictional level, but they do not link students to schools or provide achievement data. As a result, there is debate over whether survey or census data are preferable for estimating school–input effects. The estimates of effects using Census data at the aggregate level and including fixed effects tend to be greater than estimates resulting from the use of survey data at the micro level and including detailed controls but no fixed effects (Burtless, 1996). Although estimated effects are greater with Census data, Hanushek, Rivkin, and Taylor (1996) point out that the relative benefit of aggregate data as opposed to micro data is not clear. There may be more omitted–variables bias at the local level as families sort among schools and children sort across classrooms within schools, or at the aggregate level as some omitted characteristic of the region jointly determines education policy and student outcomes. Loeb and Page (2000), in a study of the effect of teacher wages on educational attainment, suggest that the ability of fixed effects to adjust for unobservables may give aggregate data an advantage over surveys. Some analyses may benefit from Census data more than from single–cohort survey data, al-

though again the available measures of student outcomes and school characteristics in Census data are extremely limited relative to the surveys.

Both national surveys and the Census data have an advantage over more geographically restricted datasets. In the United States, education policy often is set at the state level, resulting in substantial policy variation across states. National data, unlike city or state administrative data, allow cross–state comparisons, which can be used to estimate policy impacts. For example, Hoxby's (1996) paper on teachers' unions uses Census data over three decades in combination with other national data on unions to look at productivity before and after unionization. Her panel covers 10,509 districts over three decades and shows that unionization increases inputs but reduces productivity. Such an analysis would be difficult with local data that rarely span as long a time period and are unlikely to have as much variation in unionization. However, because local datasets follow individual students over time and link students to schools, they provide information and allow for empirical approaches that national datasets do not.

State and City Administrative Data

As discussed above, Census data do not have achievement measures and do not allow researchers to follow people over time or to link them to the school they attended. National survey data follow only one cohort and do not have good coverage of any single geographic region. In order to overcome these shortcomings, researchers have been assembling state–specific and city–specific datasets that link students to schools, include multiple cohorts, and include all, or at least most, students and schools in the region. These datasets provide far more thorough descriptions of schools, students, and teachers than were available previously. Researchers are taking advantage of the availability of this data by using fixed–effect techniques to reduce omitted–variables bias and incorporating more policy detail into their studies. Chicago, New York, and Texas have the most utilized administrative datasets, while Arizona, Florida, and North Carolina provide thorough data as well.[3]

The database on the Chicago Public Schools (CPS) includes information on students' school, home address, demographic and family background characteristics, special education / bilingual placement status, standardized test scores (the Iowa Test of Basic Skills), grade retention, and summer school attendance. Unique student identification numbers allow researchers to track students across years as long as they remain in the Chicago Public School System. Researchers also have access to CPS personnel and budget files, providing information on financial resources and teacher characteristics in each school and aggregate information on school population, including daily attendance rates, student mobility rates, and racial and poverty composition. Jacob and Lefgren (2002) use this data to examine the causal impact of remedial education on student achievement, focusing on the exogenous variation created by Chicago's Social Promotion Policy decision rule. Under this policy, students are required to meet specific standards in reading and math in order to be promoted to the next grade. If they do not meet the requirements in June, they must attend a six–week summer school session, after which they can retake the exams. Jacob and Lefgren create an *ex ante* quasi–experi-

[3] In Arizona, see Solmon, Paark, and Garcia's (2001) paper on charter school effects. In Florida, see, for example, Figlio and Getzler's (2002) work on accountability and disability classification. In North Carolina, see Goldhaber, Perry, and Anthony's (2002) work on National Board certification of teachers and Clotfelter, Ladd, and Vigdor's (2002) work on segregation.

ment using a regression discontinuity design to identify the impacts of these programs, comparing students who scored just below and just above the promotional cutoff. They find positive effects of summer school for third graders but not for sixth graders. Because CPS data follow all students over time, the researchers are able to compare across a narrow range of test performance. Without such thorough data, this quasi–experiment would not have been possible.

Boyd, Lankford, Loeb, and Wyckoff (2003a, 2003b, 2003c) have compiled a dataset that traces all teachers and administrators in New York State since 1969. While this dataset does not have student–specific data, it does include teacher–specific data, following teachers through classrooms and including test scores and other background information. The researchers use these data to look at the distribution of teachers across schools, the distance teachers travel from where they went to high school to their first teaching job, typical career paths, teachers' responses to the implementation of mandatory tests, and compensating differentials. This research documents the substantial differences in the qualifications of teachers across schools. Low–performing, poor, and non–white students, especially those in urban schools, are far more likely to be taught by first–year teachers, by uncertified teachers, and by teachers with low scores on teacher certification exams. These differences are clear within districts as well as across district boundaries (Lankford, Loeb, and Wyckoff, 2002). Much of these disparities result from the initial match of teachers to schools in their first teaching job, although quits and transfers exacerbate the disparities. The detailed administrative data allow the researchers to paint a more exact picture of schools, districts and teacher labor markets than was possible with data that had less coverage of schools, fewer measures of teacher and school characteristics, and did not follow individuals over time.[4]

The Texas administrative data are the most widely used of the local administrative datasets. The University of Texas at Dallas (UTD) Texas Schools Project Dataset includes individual student characteristics and test scores based on the Texas Assessment of Academic Skills Test. The dataset begins with the 1992 third grade student cohort and tracks individual students in each cohort after this time. As a result, the UTD Dataset includes over 200,000 students in each cohort in over 3,000 public schools. Hanushek, Kain, and Rivkin use these data in a series of papers looking at teachers' contributions to student achievement, peer group effects, charter school effects, the cost of switching schools, and school competition, among other topics. These papers have been characterized by the use of fixed effects to separate school, grade, classroom, and student effects. In Rivkin, Hanushek and Kain (2002), the authors put the Texas data to use with a fixed–effects model that controls for student, school–by–grade and school–by–year effects to relate differences in achievement gains between grades and cohorts to differences in school characteristics and teachers. They find that teachers have powerful effects on mathematics achievement, although little of the variation in teacher quality can be explained by the few observable characteristics of teachers. The authors find that high quality teachers are capable of erasing deficits associated with family differences in income.

Data that track individuals over time and link them to their schools and districts provide advantages over the school–level and district–level data that were available previously, and that are still used for

[4] While student–specific achievement data are not available at the state level in New York, data are available for New York City. Schwartz, Stiefel, and Kim (forthcoming) use these data and find a small positive impact of a whole–school reform initiative on student performance.

analyses in jurisdictions without such complete information. For example, the Michigan Department of Education reports whether each student scores "satisfactory," "moderate" or "low" on the Michigan Educational Assessment Program exam. It links students to schools and reports schools' racial and poverty composition, enrollment, and pupil–teacher ratios. However, the data do not track students over time. Bettinger (1999) uses these data to assess the impact of charter schools on students and public schools in the area, but is not able to account for selection into charters as effectively as Hanushek, Kain, and Rivkin's (2002) study of Texas charter schools. Hanushek, Kain, and Rivkin compare the achievement of students while they attend charter schools with the gains of the *same* students while they attend traditional public schools. Thus, the approach eliminates the omitted–variables bias associated with comparing students in charter schools with *different* students in traditional public schools.[5] They find that, on average, students perform no better and sometimes worse while in charter schools. The breadth of the data allows them to estimate average student gains for each school, showing that the distribution of charter and public schools is similar except that some charter schools perform particularly poorly. Neither the estimation technique nor the picture of the variation across schools in their value–added would have been feasible without student–level longitudinal data and complete coverage of schools and students.

While the analyses of these datasets advance education policy research, they do have limitations. As with most studies of education policy, the trick is establishing causality. The most common approach used with these administrative datasets is fixed–effects modeling. This technique allows researchers to remove many of the unobservables. However, fixed–effects methods are not a cure–all solution. When a series of fixed effects are included in a model, the identification comes from a subset of the total variation. It is important to understand where the identifying variation is coming from. For example, if we are interested in the impact of teacher characteristics on student outcomes, it might not be appropriate to identify our estimates from differences in the characteristics of teachers within the same school. Suppose that we find within schools (including school fixed effects) that teachers with higher test scores do not help their students to achieve more than teachers with low test scores. This may show that, on average, higher–scoring teachers are not more productive than lower–scoring teachers. Alternatively, hiring authorities may balance test scores with other attributes that we do not observe. Thus a teacher with a lower score would likely be "better" on the unobserved attribute. While the inclusion of school fixed effects in such a model rids the estimation of unwanted biases due to differences across schools that affect student achievement, it also eliminates the variation that we may need to accurately estimate the effects. Similarly, fixed effects may restrict the identifying sample to a specific group. If this is the case, it is useful to consider whether the effects for this sample are likely to differ from those of the omitted group. For example, some analyses may identify effects from students who move and thus experience two different policy regimes. However, students who move may differ from those who do not, and moving itself may impact their outcomes, therefore causing bias or reducing the generalizability of the results.

[5] Two potential biases remain. First, there may be omitted time–varying characteristics of students who are associated with both charter school attendance and achievement gains. Second, students who select into charter schools may be particularly prone to benefit from these schools, and thus, any estimates of effects would not be generalizable to the full student population.

Because of the aforementioned problems with using fixed effects to establish causality, researchers are also using an instrumental variable approach to estimate policy effects with longitudinal administrative datasets. Hoxby's (2002) research on the effects of reduced class size uses natural variation in enrollment that trigger the implementation of minimum or maximum class size rules. With this instrumental variable in use to isolate the effect of class size, she uses the Connecticut School Districts longitudinal dataset (which consists of 649 elementary schools that belong to 146 elementary school districts) to compare the class size and achievement of adjacent cohorts who immediately precede and succeed each triggered change. Angrist and Lavy (1999) use a similar methodology and instrumental variable with data from Israel. The greatest difficulty with instrumental variables is finding an instrument that is suitably correlated with the policy variable, but not correlated with the error term (Bound, Jaeger, and Baker, 1995). Another difficulty of instrumental variables is similar to, and often far worse than the drawback discussed above for fixed–effects approaches. The instrumental variable restricts the variation identifying the effects. As an example, in both Hoxby's and Angrist and Lavy's work, the variation comes from unanticipated differences in class size. Does this differ from the anticipated or longer–term differences, due, for example, to a class size reduction policy (in a state with a tradition of not changing policies often)? How big is the variation in class size created by these instruments? Given the diversity of findings, it is useful to identify the variation in order to draw policy implications from the results.

Some researchers have used matched samples of schools in combination with local administrative data in order to estimate policy effects. Matched samples are beneficial when the treatment groups differ on pre–treatment characteristics and either the treatment has a differential effect across groups or the outcome trajectories vary across groups. Finding an appropriately matched sample is not always an easy task. For example, when considering the impact of charter schools, it may not be enough to match by background characteristics or prior achievement, since selection of alternative schools already signifies a difference between the treatment and control group. However, in cases in which selection is not a large concern, matched samples may provide a useful tool. Angrist and Lavy (2001) estimate the effect of in–service teacher training on achievement in Israeli elementary schools using a matched–comparison design. They use a student fixed–effects technique to examine how the Thirty Towns program, which increased resources for teacher training, improved student achievement on standardized tests. The authors compare a population of fourth–grade pupils enrolled in "treated schools" in 1994 with a population of fourth–graders who attended a sample of "control schools," selected for their comparability and for data collection reasons. Their use of "controls" and "treatments" is made possible through an extensive student–level dataset, the ability to look at schools as they progressed before and after the Thirty Towns "treatment," and the fact that schools did not self–select into the program. They find that increased funding for teacher training led to considerable and statistically significant increases in math and reading test scores relative to control schools that received no extra funds for training. Their preliminary cost–benefit analyses imply that teacher training may be as cost–effective as lengthening the school day and cheaper than reducing class size to improve student achievement.

Establishing causal effects can be very difficult. One of the impediments comes from the policy environment, in which

many changes are simultaneously occurring. In his paper on accountability, Jacob (2002) notes that there were other policies and / or programs in place in Chicago (and nationally) at the time of his study on accountability effects that may have impacted overall performance. He cannot control for all of these possibilities. For simultaneity reasons such as the one encountered by Jacob, it can be difficult to prove causality when using local administrative data, no matter how thorough the dataset.

Accountability itself poses problems for using administrative datasets to assess student achievement. The achievement measures used in national surveys do not have awards or sanctions tied to them. Because of this, schools are unlikely to be working towards goals particular to the tests given. However, as states implement accountability policies there is the potential for time–series analysis to be skewed as teachers and school administrators begin "teaching to the test," raising achievement scores in the areas being tested in standardized assessments, but possibly lowering achievement in other areas that are not tested. When the measures of achievement used for research on predictors of student achievement are the same as those used to assess schools for accountability, the estimates may be biased. For example, a study might find that low–income students have as "effective" teachers as higher–income students, when in reality, the similarity may be a result of an increased focus on test coverage for low–income students given the incentives created by the accountability system. Higher–income students may be seeing similar or even smaller gains on these tests, but may be gaining elsewhere.

A final drawback of state and city administrative data is their limited geographic scope, combined with large variation in policy environments across jurisdictions, which makes generalizing results difficult. Boyd, Lankford, Loeb, and Wyckoff (2003c) find that teachers in New York State are likely to teach very close to where they grew up. Sixty–one percent of teachers in their sample take their first teaching job within 15 miles of the district from which they graduated; 85 percent, within 40 miles. While other states may have teachers working in similarly close proximity to home, states with greater immigration and faster growing student populations, such as California and Texas, may differ.

While generalizing results to other jurisdictions is difficult, the limited geographic scope of local data has led researchers to account more fully for the details of education policies. Many policies are tricky to understand, often containing seemingly peripheral elements that affect the impact of the reform. Researchers looking across states tend to simplify the reforms by comparing states with and without a general type of policy. Murray, Evans, and Schwab (1998), for example, compare states with and without court–mandated school finance reforms. While they find that states with such reform have equalized more than other states, they lose much of the variation due to the very different policies implemented as a result of litigation. Researchers using state and city datasets have tended to consider the reforms more closely. By doing this, they often are able to make use of the variation in policy incentives across local jurisdictions to create quasi–experiments for estimating policy effects. Cullen (forthcoming), for example, examines the effect of fiscal incentives on student disability rates. Using Texas data, she compares the changes over time across districts differentially affected by the legislative change in funding formulas for special education students. In another example, Cullen, Jacob, and Levitt (2002) use lotteries for intra–district school choice in Chicago to create a quasi–experiment for assessing the impact of choice on student outcomes.

Although the new state and city datasets clearly have their drawbacks, we should not understate their advantages. As noted above, individual fixed effects have reduced, though not eliminated, concerns of omitted–variables bias. Such data bring more detailed information on students, teachers, and schools than we have had before, except in surveys with only light coverage of most geographical areas. The thorough coverage of this data allows us to look much more closely at differences across districts within states, schools within districts, and students and teachers within schools. Similarly, with these data we can better identify not just average effects but the distribution of effects and the differential impact of policies across groups and under disparate conditions.

Experimental Data

Experiments randomly assign participants to treatment and control groups and thus remove the worry that selection into the treatment group is biasing estimates of the effect of the treatment. Two sets of experiments have gained particular attention, one addressing class size reduction and the other, vouchers.

In the Tennessee Student/Teacher Achievement Ratio (STAR) experiment, over 6,000 students in 79 schools across urban, suburban, and rural areas were randomly assigned to small and large kindergarten classes, for the most part continuing the assignments through the third grade. The project was funded by the Tennessee legislature at a cost of $12 million over four years. Children who came into experimental schools mid–experiment were randomly assigned to small or large (or large with aide) classes and added to the experiment. Participants who skipped a grade or left the school were taken out of the sample. Researchers agree that the STAR experiment demonstrates positive effects of smaller classes in the kindergarten year, particularly for minority and low–income students, although no further gains are evident in the later elementary years (Krueger, 1997; Hanushek, 1999). While there were no apparent effects of smaller classes after kindergarten in elementary school, Krueger and Whitmore (1999) find class size does affect students long after they complete elementary school. Their results show that students who attended a small kindergarten class as part of the STAR experiment were more likely to take the ACT and SAT exams at the end of high school.

Voucher experiments have been carried out in Dayton, New York and Washington D.C.[6] A voucher program in Milwaukee also provides random–assignment data. The Milwaukee Voucher Program was implemented in 1990–91. It gave children from families at or below 175 percent of the poverty level the option of using public money to enroll in private schools. Students who applied were randomly selected in each year for a slot in a specific grade and school. Initially the vouchers were restricted to 1,500 students, but this limit was not binding during the first years of the program. The program was restricted to non–religious private schools and schools were not allowed to charge voucher students supplemental tuition. In 1997 the program expanded to allow for 15,000 students and to include religious schools, providing students with a voucher approximately equal to the per–student cost in Milwaukee public schools, $5,500 in 1997 (Carnoy, 2001). The Milwaukee program was not designed as an experiment. However, because not all applicants were able to find slots in partici-

[6] A voucher experiment has begun recently in Charlotte, North Carolina, as well. A voucher program in Cleveland has been operating since1996–97, but the data available are non–experimental. A voucher program in San Antonio, Texas, begun in 1998–99, and a scholarship program in the San Francisco Bay Area offer funds for students to attend private schools but are not experiments.

pating private schools, the remaining students formed a natural control group: as noted below, missing data make assessment of the treatment effect difficult.

While the Milwaukee program was publicly run, the voucher programs in New York, Dayton, and Washington were privately sponsored and designed specifically as experiments. The School Choice Scholarships Foundation in New York City started in the fall of 1997 and provided 1,300 scholarships to students who were entering first through fifth grade, were currently attending public school, and qualified for free lunch. The vouchers were worth $1,400 per year for three years, to be used at secular or religious private schools (Krueger and Zhu, 2002). The Parents Advancing Choice in Education Program in Dayton Ohio began in 1998 with 515 students who had been enrolled in public school and 250 who were already attending private school. In 1999 the maximum scholarship was $1,700 for elementary school students and $2,300 for high school students. The Washington Scholarship Fund, begun in 1993, increased its funding in the late 1990s so that by spring of 1998 1,000 students were offered scholarships. Students with family income below the poverty line received the lesser of $1,700 or 60 percent of tuition (Wolf, Peterson, and West, 2001). All three of these programs required students' families to pay part of the tuition for their religious or secular private school. Over 20,000 students filled out initial applications for vouchers in New York City, over 3,000 applied in Dayton, and over 7,500 applied in Washington (1,582 of whom met the criteria for inclusion). Voucher recipients were chosen by lottery and experiments were set up to compare the outcomes for students who won the lottery with other applicants (Howell, Wolf, Peterson, and Campbell, 2000).[7]

While experiments have been touted as the best form of research on education policy, it appears to be difficult to implement a true experiment in education. Different researchers assessing the same experimental data do not always, or even often, agree on the policy effects. A number of problems have emerged in these and other experiments. First, many experiments have non–random attrition. Families move in and out of school districts, and parents who are unhappy with their children's situations move their children into different schools. Attrition is a serious enough problem in most education experiments that it is difficult to sustain an experimental design for more than a couple of years.

A second limitation of experiments is that they tend only to be able to estimate the impact of the treatment on a restricted group of students. This is not always a sign of bias, but in some cases it is an indication of the need to acknowledge the limited group to which the results apply. The voucher experiments have focused on low–income students. This is a restricted group, but one in which policy makers may be particularly interested. However, consider experiments in which families apply for vouchers and are then placed in either a treatment (received a voucher) or a control (did not receive a voucher) group. The act of applying indicates that these families differ from those who did not apply, making it difficult to generalize the results to the population that did not apply. If those who apply are those likely to gain the most from private schooling, then the results may over–estimate the treatment effect for the full population. The experiment in this case cannot give the causal effect for the larger group.

In addition to these difficulties there are the typical problems associated with experiments. For example, some experimen-

[7] See the third section for a brief summary of the results of the voucher experiments. Gill et al. (2001) provides a more thorough review of the studies assessing voucher effects.

tal subjects (teachers, administrators, parents, and even students) may temporarily change their behavior when they know that they are being observed or evaluated, making policies appear to have productivity effects that would not be as strong if the policy were implemented fully.[8] Additionally, administrators may purposely skew the treatment groups because of pressure from parents to ensure that their children are receiving the beneficial "treatment." These types of problems may be particularly severe in education experiments since it is necessary to have the consent of all of those involved. Moreover, there is the risk that some participants have incentives to perform differently not only because they are being observed, but also to influence policy decisions. For example, if teachers would like their state to implement class size reduction and the experiment will influence policy decisions, then those in smaller classes have incentives to perform better than those in larger classes.

Finally, experiments are small in scale, rarely generating the indirect effects observed with full–scale policy implementation. A small voucher program will not generate the increased demand for private school spaces likely to result from a bigger program. Consequently, it will not allow researchers to estimate the elasticity of supply of private schools (either quantity or quality). Given that private schools are heterogeneous, these experiments cannot show whether vouchers will have the same effect when demand increases private school enrollment. Similarly a class size experiment is unlikely to increase the demand for teachers as much as a state level class size reduction policy. As such, the experiment will not allow the researchers to assess the impact of class size reduction on the teacher labor market. When California reduced class size in the early elementary grades, teachers transferred to higher–performing schools, leaving the poorest schools with vacancies and inexperienced teachers (Wexler et al., 1998). Using data from California, Jepsen and Rivkin (2002) find that a reduction in class size of ten students (the goal of the policy) increased the number of students meeting the national average in math by 4 percentage points and by 3 percentage points in reading. However, on average, a first–year teacher reduced this gain by 3 percentage points in both reading and math. It is likely that some students were worse off with the implementation of class size reduction due to teacher labor market dynamics.

The debates over both the STAR experiment and the voucher programs have received substantial attention, and researchers have used a number of econometric techniques to overcome flaws in experiments and assess the treatment effects. Greene, Peterson, and Du (1997), in their evaluation of the Milwaukee program, compare students who received vouchers to students who applied and did not receive vouchers. Because the applicants were randomly chosen for vouchers, this comparison is a sensible one for assessing treatment effects. However, over half of the unsuccessful applicants for vouchers never returned to the public schools and thus were lost from the sample. Those who did return were from less–educated, lower–income families. Because of this, the control group with data no longer reflected the complete control group or the larger population. Witte, Sterr, and Thorn (1995), taking an approach used commonly with non–experimental data, compare a sample of public school students with students who received vouchers. They attempt to control for students' prior test scores in their analysis, using a value–added approach to individual student achievement, but gain no benefit from the experimental as-

[8] This is known as the Hawthorne Effect and was first noted in the Western Electric Hawthorne plant.

pect of the lotteries. Rouse (1998) uses both of the comparison groups—public school students and non–accepted applicants—and finds no effect of vouchers on reading scores but a positive effect of 1 to 2 percentage points per year on math scores.

Experiments have the potential to substantially increase our understanding of school policy effects. However, in order to elicit the benefits we need to pay close attention to implementation. Even in the best of circumstances the results of experiments need to be placed in a structure that enables researchers to take the results and incorporate the indirect effects of larger scale policies, considering how representative the treated group in the experiment is of the larger student population.

FINDINGS

The empirical literature by economists addressing school policies is huge and growing. Although we limit ourselves to three current issues in school reform policy, we are only able to include a fraction of the papers on these topics. As such, we are presenting a necessarily incomplete review, with specific papers selected to point out important data issues and to address contemporary education policy concerns. What follows is a summary of some interesting findings.

School Finance Equalization

The past 30 years have seen a dramatic shift in school finance from the local level to the state level. Some of the shift has been driven by property tax reform (i.e., Michigan) but much has been the result of litigation for inequitable or inadequate schooling (i.e, California). Early studies use data on California, the first state to centralize school finance as a result of litigation, to assess the impact of finance reform. They find a decrease both in spending variation and average spending levels as the result of reform (Downes, 1992; Sonstelie, Brunner, and Ardon, 2000). As more states faced court challenges and centralized finance, cross–state comparisons became viable. Murray, Evans, and Schwab (1998) use Census of Governments data at the district level and a model that incorporates state fixed–effects to examine how reform altered the time path of within–state inequality. They find that court–ordered finance reform reduced within–state inequality in spending by 19 to 34 percent, and that successful litigation against unequal funding formulas reduced inequality by raising spending in the poorest districts while leaving spending in the richest districts unchanged, thereby increasing aggregate spending on education.

Why did the results of the cross–state analysis differ from the California results? Foremost, finance reform in California placed much greater restrictions on local revenue raising than did reforms in other states (Hoxby, forthcoming; Loeb, 2001). In addition, substantial demographic changes in California during the period of reform confound studies that use before–and–after comparison (Downes, 1992). Hoxby (forthcoming) is able to reconcile some of the differences in the impact of reform across states by incorporating specifics of the reform, including the tax price of education spending, into a national study. She finds that some equalization, but not radical equalization as implemented in California, can result in the leveling up that Murray, Evans, and Schwab (1998) observe. In–depth, state–specific case studies of finance reform—such as those of Kansas (Duncombe and Johnston, forthcoming), Kentucky (Flanagan and Murray, forthcoming), Michigan (Cullen and Loeb, forthcoming), Texas (Imazeki and Reschovsky, forthcoming), and Vermont (Downes, forthcoming)—also shed light on the reasons why the impact of reforms varies across states. They describe the reforms in substantially more detail than cross–state

analyses and directly identify tensions created, for example, by disparities between district demand for education spending (how much they would like to spend) and their allowed spending under the reforms.

While it is clear that school finance reform directly affected operating expenditures per pupil across districts, the effect on student outcomes is less obvious. Card and Payne (1998), using Census of Governments data, find some evidence that the reforms decreased the gap in SAT scores between students from different economic backgrounds. However, the SAT is not taken by a random sample of students. While the researchers can adjust their estimates for the percent of students in each state taking the SAT, they cannot assess the impact of finance reform on the portion of the student population that is unlikely to take the SAT. Given that the focus of much of the litigation has been on the lower end of the distribution, their analysis may not address the part of the achievement distribution that is most relevant for assessing reform effects. This issue highlights one of the main disadvantages of using national census data: the lack of student outcome measures. Downes and Figlio (1999) take an alternative approach, linking the Study of the High School Class of 1972 with NELS88. By using survey data, they are able to assess achievement effects across the distribution, finding small positive effects of school finance reform overall, but some increase in the variation of scores, though the results are not robust to alternative specifications.

A benefit of these national analyses is that they can compare across states. Within–state analyses tend to be limited to the before–and–after approach, though they can also compare across local jurisdictions that are differentially affected by the school finance reform. Downes (1992) and Cullen and Loeb (forthcoming) use state–specific data and find no effect of finance reform on student performance in either California or Michigan, two states that saw substantial change in their funding systems.

A large body of research has addressed responses to school finance equalization. In general, studies have found that districts and schools are quite responsive to incentives created by finance systems. Many, though not all, of these studies use state administrative data. While these data often allow for a detailed assessment of policy change, the results may not be generalizable to other states. Cullen (forthcoming), for example, finds that changes to the funding formula for special education students in Texas can explain over 35 percent of the growth in Texas student disability rates over the past two decades. Brunner and Sonstelie (1996) find that parents also react to finance reform. Private contributions to public schools grew substantially in California after litigation spurred legislation that severely constrained districts from raising local tax money for schools. Similarly, private school attendance rose following the reforms in California (Downes and Schoeman, 1998).

Another indication that voters respond to finance reform comes from the capitalization of reform in housing values. Guilfoyle (1998), for example, using Michigan individual home sale data and the changes that resulted from school finance reform, finds that a $1 tax differential leads to a $5.20 home value differential. He also finds significant effects of spending ($100 increase in per pupil spending would raise home values 0.4 to 0.6 percent). These estimates imply that if a community raised spending through property taxation, the effects would come close to canceling one another. In a cross–state comparison of district–level census data, Dee (2000) addresses the same question and finds similar results. Districts with gains in education spending due to reform experienced differentially greater

growth in property values. These results are consistent with Black (1999), who compares prices of houses near the boundaries of Massachusetts' school districts and finds that a 5 percent increase in test scores results in a 2.1 percent increase in housing prices.

Three other findings are also worth noting because they have implications for modeling the impact of school finance reforms. First, there is income heterogeneity within school districts. This effect is large enough that the median voter in some states has a lower tax price for education spending under local control than with state funding, even though the income distribution within each state is skewed (Loeb, 2001). Models of school finance reform often assume a homogeneity of income within districts, an assumption that may incorrectly influence conclusions. Second, the link between residents' income and district tax base wealth can be small due to non–residential property in a district. The strong negative relationship between property wealth and per pupil spending that led to California's reform coexisted with a much more even distribution with respect to income (Sonstelie et al., 2000). This disparity between income and property wealth provides an additional explanation for the weak relationship between finance reform and student achievement. Urban areas often house low–income students with the lowest achievement levels, yet also have substantial non–residential property. If finance reforms do not improve the resources in these low performing schools then they are unlikely to affect the distribution of student achievement. Finally, there are large differences across districts in the cost of providing a similar education, but cost differences are difficult to separate from differences in the efficiency of production (Duncombe and Yinger, 1999). A better understanding of the contribution of school funding to student outcomes requires a better understanding of what contributes to the cost difference. Again, data that provide more detail on schools and students give researchers an opportunity to improve this understanding.

Accountability

As states have assumed more responsibility for funding schools, they have also increased their monitoring of schools. On January 8, 2002, President Bush signed the "No Child Left Behind" Education Bill, requiring states to adopt standardized testing for students in grades three to eight, and to use the test scores to grade schools. Even before this bill, "high stakes" standardized testing had been playing an increasing role in states' public education systems. By 2000, 28 states had passed legislation to establish minimum test standards required for a student to graduate from high school. Some states also use test scores to determine grade promotions and summer–school enrollments. Most states publish student test score information by school or district, and some use these scores as a basis for rewards or interventions. Currently, at least 35 states use student test scores to determine school ratings or school accreditation status. Of these states, 14 use student performance measures to assign discrete grades or ratings to all schools and / or school districts.

There is no experimental data to assist researchers in their attempts to determine the effects of accountability on student achievement. It is difficult to imagine an experiment that could do this well. As such, studies use either national or state–specific longitudinal datasets to link accountability programs and student outcomes. Carnoy and Loeb (2003), using state–level data on student outcomes, find evidence of increased achievement linked to accountability. They use the National Assessment of Educational Progress (NAEP) for fourth and eighth graders, a

low–stakes exam not linked to any accountability system.[9] They rank the strength of state accountability systems and then use state–level data to look at changes between 1996 and 2000 on the NAEP math exam. They find that states with stronger accountability systems show greater growth in eighth grade test scores for all racial/ethnic groups, both at the basic level and at the proficient level. They also find increased scores for Hispanic and black students in the fourth grade.

A clear disadvantage of Carnoy and Loeb's approach is that they do not identify the impact of any single aspect of the accountability system, but rather focus on the general intensity of incentives created by the complete accountability program. Other studies have looked at specific aspects of accountability or testing. Frederikson (1994) uses a similar approach and NAEP data to look at the effect of the introduction of minimum competency tests in the early 1980s. He finds that states that implemented these tests show a greater increase in math scores. However, Jacob (forthcoming), using the NELS88 national survey data, does not find a difference across states in student learning associated with minimum competency testing. The advantage of the Jacob study is his use of student–level data that allows the researcher to estimate test–score gains and adjust for differences in students across policy regimes. The disadvantage is that the survey only gives a glimpse of one moment in time and thus Jacob is not able to compare one policy environment to another within the same state.

Because programs vary across but not within states, national data provide an opportunity to look across jurisdictions with different policies. However, state–level analyses have, at most, 50 observations. Accountability systems are quite different from state to state. Even those with the same score on the index used by Carnoy and Loeb may have quite different systems that create disparate incentives for districts, schools, teachers, and students. It is difficult to capture the variation in policy and control for necessary differences across states with so few degrees of freedom. Because of this, it is worth looking at each state separately and in more detail, although studies that do this are generally restricted to before–and–after analyses in which it is difficult to show causality since other factors have changed simultaneously. Jacob (2002) finds large gains in math and reading achievement in Chicago following the introduction of accountability, but determines that these gains are largely the result of improved test–taking skills and increased student effort, since the same gains are not evident on a low–stakes exam. The difference between the Chicago results and the national comparison in Carnoy and Loeb (2003) may stem from the particulars of the Chicago program. In the first years of the test, there was little monitoring of teachers. In fact, teachers "cleaned" the tests of their students to make the scantrons easier to read, providing opportunities for teachers to cheat (Jacob and Levitt, 2003).

Local administrative data, when combined with the specifics of the state accountability system, can provide opportunities for quasi–experiments. In New York, for example, Boyd, Lankford, Loeb, and Wyckoff (2003a) are able to compare the behavior of fourth grade teachers to the response of other elementary school teachers, because the mandatory tests were given only in the fourth and eighth grades. They find that new teachers are much less likely to be placed in the tested grade, indicating a responsiveness of schools to the created incentives.

[9] As discussed earlier, research on accountability systems that uses the same tests used by the accountability programs to judge student outcomes can be skewed.

The stated objective of higher standards (for example, requiring all students to pass ninth grade algebra or biology to graduate from high school) and statewide assessment, including high school exit exams, is to increase schools' focus on the amount students learn. This new focus, however, may have unintended consequences, at least in the short–run. Raising the bar on student learning in high school may make it more difficult for students to pass courses, hence increasing student retention and decreasing graduation rates. Dee (2002) uses 1990 Census data and a comparison across cohorts to analyze the introduction of minimum competence testing and course graduation requirements. He finds reductions in educational attainment, particularly for black students. Jacob (forthcoming), using NELS data, finds similar results for low–achieving students. Carnoy and Loeb (2003) find only slight evidence that the current accountability systems affect student progression through high school, but their accountability measure focuses equally on elementary and secondary school accountability and thus may be a poor measure for assessing the particular affect of graduation requirements.

Researchers find strategic responses by schools and districts to accountability. In general, the local administrative datasets provide the necessary detail and coverage for these analyses. As noted above, Jacob and Levitt (2003) find strong evidence that teachers in the Chicago schools cheat on the exams to improve the scores of the students in their classes. Cullen and Reback (2002) use data on Texas and find that schools reclassify low–income and low–performing students as special education in response to accountability. Figlio and Getzler (2002) find similar results using Florida data. Jacob (2002) finds that teachers responded to the new policy by increasing special education placements, retaining students, and decreasing emphasis on low–stakes subjects that would not be tested. Overall, while accountability may improve student achievement, research indicates that it elicits a variety of responses which may detrimentally affect students. As in the school finance literature, national datasets give a broad view of average effects across states, while local datasets provide more detail on the impact of specific reforms.

Competition

Mandatory testing is one approach to monitoring schools, but competition provides a second, decentralized option. When parents can choose the schools for their children, they create market pressure for schools to use their resources effectively. One benefit of competition over direct accountability is that schools provide multiple outcomes, many of which are difficult to measure. Hoxby's (2000b) work, using NLSY and NELS88 in combination with census data, has been particularly influential in suggesting that competition can improve school effectiveness. She uses variation in district size due to naturally occurring rivers and streams to compare the achievement in metropolitan areas with varying numbers of school districts, and finds that the eighth and twelfth grade reading achievement and tenth grade math achievement of non–minority students is higher in metropolitan areas with more districts. In addition, spending in highly competitive metropolitan areas is lower than in less–competitive areas, while income and grade attainment is higher. State–specific administrative data can also provide evidence on competition. Hanushek and Rivkin (2003) use the Texas data and a competition measure based on the extent to which students within each metropolitan area concentrate in large districts (the Herfindahl index) to find that increased competition improves teacher quality.

Another approach to testing the potential impact of market–induced competition has been public–private school com-

parisons. While there are numerous studies in this vein, the selection of students into school type makes identifying causal effects difficult. Most studies have not found a significant difference in value–added to achievement between public and private schools, though Neal's (1997) research using NLSY indicates that Catholic school students, especially minority students in urban environments, are more likely to enroll in college than their public school peers. A number of recent policies aim to increase competition within the public sector, including charter schools, vouchers, and both intra– and inter–district choice among traditional public schools:

Charter Schools

Charter schooling is a recent phenomenon, dating back to Minnesota in the 1992–93 academic year. Currently, 39 states, the District of Columbia, and Puerto Rico have passed charter school laws (Iowa, New Hampshire, and Tennessee have charter laws but no charter schools yet). Approximately 1 percent of students attend an estimated 1,800 charter schools across the United States (http://www.uscharterschools.org). While the number of charter schools has been growing rapidly, the changes are recent enough to make assessment difficult.

Studies of charter schools to date use state administrative data. National data are not current enough to address this recent phenomenon and there are no relevant experiments. This lack of variety in data, combined with the newness of charter schools and the impact on children of switching schools, confounds the estimates of charter school effects. As a result, the evidence on current charter schools may not be generalizable to the effect of charter schools in the long run. Thus far, there is little evidence that charter schools improve student learning. However, there is also little evidence that charter schools "cream–skim" the most able students from tradition public schools, an initial concern.

Eberts and Hollenbeck (2001) and Bettinger (1999), in their studies of charter schools in Michigan, find that charter school students are gaining less on the state test than are students in traditional public schools. However, as discussed above, Michigan data do not allow researchers to follow students across schools, thus the results may be biased by student selection into schools. Hanushek, Kain, and Rivkin (2002) use the more complete Texas data, but similarly find that, on average, charter schools have a negative effect on achievement. Hanushek, Kain, and Rivkin also estimate the average value–added of all Texas schools and find that charter schools and traditional public school are similar through much of the distribution but that charter schools have a longer left tail, indicating that some of the charter schools are doing a particularly poor job of adding to students' learning. Gronberg and Jansen (2001), also looking at Texas, find a positive effect of charter schools for "at–risk" students, but a negative effect for other students.

The impact of charter schools appears to depend on the length of time they have been in operation as well as the students they serve. Eberts and Hollenbeck (2001) distinguish schools by their years of operation and find that charter schools that have operated for longer periods of time have less of a negative impact on students' achievement relative to traditional public schools. Hanushek, Kain, and Rivkin's (2002) negative results for mathematics is driven entirely by the first year schools in their sample and the negative reading effect is driven entirely by first and second year schools. Studies using Arizona data echo these results. Solmon, Paark, and Garcia (2001) use a similar approach to the Texas studies and find a significant negative impact of charter schools in the first year, but significant positive effects in the second and third years. Given the substan-

tial differences in charter school policies across states, the consistency of the findings is surprising. Charter schools do not appear to increase student achievement in the first years of operation and may, in fact, have a negative impact. Evidence suggests that this negative impact decreases over time, with charter school tenure.

An initial concern with charter schools was that they would "cream–skim" students, taking the highest achieving students out of traditional public schools. This does not seem to be the case. Eberts and Hollenbeck (2001) find that in the first few years of implementation, Michigan charter school students were more likely to be eligible for free or reduced price lunch than students in traditional schools in the surrounding districts. This difference has decreased over time. Similarly, Bettinger (1999) finds that Michigan's charter schools usually attract students who are performing poorly relative to neighboring public schools. In Arizona, Solmon, Paark, and Garcia (2001) find that charter schools tend to take students with lower than average scores, though they are also more likely to be white and speak English as their primary language.

It is also worth noting that test scores may not be a useful measure of charter school performance. Many charter schools have non–academic themes or concentrations that do not translate into higher test scores, although they may be achieving their own stated goals. Parents who select charter schools may be more satisfied with charter schools than with their children's traditional public schools; although, since those parents chose to leave their traditional school, they do not represent the general population. There is some evidence that charter schools operate more efficiently. Hoxby (2002), using self–collected survey data, finds that charter schools hire more highly qualified teachers (as measured, for example, by selectivity of their undergraduate institution or by math and science skills) and pay a higher premium for these qualifications.

Vouchers

While voucher programs are, in many cases, even more recent than charter schools, the assessment of vouchers has the advantage of experimental designs. Yet, even with experiments, the results are far from conclusive. As noted above, non–random attrition makes the Milwaukee program difficult to assess. The experiments in New York, Dayton, and Washington provide the least controversial estimates of voucher effects. Peterson, Howell, Wolf, and Campbell (2003) and Krueger and Zhu (2002) find that the New York City scholarship experiment improved math scores more than reading scores, though all positive effects were limited to African–American students. Again, even within an experiment, the results differ by approach. Krueger and Zhu, including a larger sample, found smaller effects than Peterson, et al. As in New York, the Dayton and D.C. experiments demonstrated positive effects of vouchers, but only for African–American students. In Dayton, the effects are only in reading and not in math; in D.C. the results fade out after two years (Peterson, et al., 2003). Overall, there is little evidence of positive effects for non–African–American students, but some of the results for African–American students are promising.

Angrist, Bettinger, Bloom, King, and Kremer (2001) study a voucher program in Colombia. They find that voucher "winners" (those who won a lottery to receive a voucher) scored slightly higher than "losers" (those who applied for, but did not receive a voucher), especially in reading. Lottery winners completed more schooling than losers on average, and were less likely to repeat grades. This last finding might stem from the reduced incentive for grade repetition among

voucher–accepting schools, as voucher students would loose their voucher if forced to repeat a grade.

As noted above, the estimated impact of vouchers may not generalize well to the group of families that did not apply to the program. Metcalf (1999), for example, finds that scholarship applicants had a high level of dissatisfaction with the public schools. In addition, the results of the experiments may not be generalizable to a large–scale voucher program that would rely on substantial movement in the supply of private schools.

Choice among Traditional Public Schools

Charter and voucher schools are not the only school choice options available within the public school system. Currently, one in seven school districts nationally allows students to transfer schools within the same district (National Center for Education Statistics, 1996), and nearly every major urban district has at least one magnet school that attracts students district–wide (Blank, 1990). There is little evidence on the impact of this choice. Cullen, Jacob, and Levitt (2000), using data on Chicago schools, find no benefit of choosing non–neighborhood schools for Chicago students, but they are unable to ask whether the availability of choice improves achievement overall. There is evidence, however, that choice within traditional public schools can influence housing values. Using data on the housing values before and after Minnesota implemented inter–district choice, Reback (2002) finds that property values declined substantially in districts that were net receivers of students and increased in districts that were net senders.

CONCLUSIONS

There are inherent difficulties in using empirical analyses to predict policy effects; while some datasets are clearly better than others, none provide all the answers. Census data provide national coverage and the ability to link localities over time. However, the data do not follow individuals or link individuals to the schools they attend. In addition, Census data have few measures of school characteristics or student outcomes. National survey data have the most complete measures of schools and students, but follow only one cohort and have limited coverage of localities. The recent state and city administrative datasets overcome some of these weaknesses by following multiple cohorts, linking students to schools, and including measures of student achievement. They also tend to cover all students and schools within their jurisdiction. These data are geographically limited, however, and thus not useful for assessing the impact of policies that do not vary within their borders. Moreover, researchers using these datasets or the national data have difficulty establishing causality, often due to the selection of students into schools. Experiments may overcome this, but they are difficult to implement for many education questions and they have not been large enough to elicit indirect effects.

Because of the divergence of data and the surge in research addressing school policy questions, research findings are difficult to interpret. In this paper, we do not attempt to sort out the findings, only to demonstrate that techniques and datasets have advantages and disadvantages for addressing specific research questions, and that analyses using different approaches can result in different estimates. These inconsistencies do not imply that econometric analysis of school reforms are without merit. The discrepancies are likely to be due to differences in the identifying variation, to heterogeneity of policies, to the influence of local conditions on policy effects, and to the differential effects of a single policy on heterogeneous individuals. They are the natural result of trying to use better data

and more accurate econometric techniques. Before we can draw policy implications we need to sort out the findings.

REFERENCES

Angrist, Joshua D., Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer.
"Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." NBER Working Paper No. 8343. Cambridge, MA: National Bureau of Economic Research, 2001.

Angrist, Joshua D., and Victor Lavy.
"Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114 No. 2 (May, 1999): 533–75.

Angrist, Joshua D., and Victor Lavy.
"Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools." *Journal of Labor Economics* 19 No. 2 (April, 2001): 343–69.

Bettinger, Eric.
"The Effect of Charter Schools on Charter Students and Public Schools." National Center for the Study of Privatization in Education, Occasional Paper No. 4. New York: Teachers College, Columbia University.

Betts, Julian.
"Do School Resources Matter Only for Older Workers?" *The Review of Economics and Statistics* 78 No. 4 (November, 1996): 638–52.

Betts, Julian R., and Jamie L. Shkolnik.
"The Behavioral Effects of Variations in Class Size: The Case of Math Teachers." *Educational Evaluation and Policy Analysis, 21* No. 2 (Summer, 1999): 193–213.

Black, Sandra E.
"Do Better Schools Matter? Parental Valuation of Elementary Education." *The Quarterly Journal of Economics* 114 No. 2 (May, 1999): 577–99.

Blank, Rolf K.
"Educational Effect of Magnet High Schools." In *Choice and Control in American Education vol. 2*, edited by William H. Clune and John F. Witte, 77–109. New York: The Falmer Press, 1990.

Bound, John, David A. Jaeger, and Regina M. Baker.
"Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variables is Weak." *Journal of the American Statistical Association* 90 No. 430 (June, 1995): 443–50.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff.
"Do Mandatory Tests Affect Teachers' Exit and Transfer Decisions? The Case of the 4th Grade Test in New York State." Working Paper. 2003a.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff.
"Analyzing Determinants of the Matching of Public School Teachers to Jobs." Working Paper. 2003b.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff.
"Implications of the Geography of Teacher Labor Markets for Teacher Recruitment." Working Paper. 2003c.

Brewer, Dominic J., Eric R. Eide, and Ronald G. Ehrenberg.
"Does It Pay to Attend an Elite Private College? Cross–Cohort Evidence on the Effects of College Type on Earnings." *The Journal of Human Resources* 34 No. 1 (Winter, 1999): 104–23.

Brunner, Eric, and Jon Sonstelie.
"Coping with Serrano: Voluntary Contributions to California's Local Public Schools." In *Proceedings of the 89th Annual Conference on Taxation*, 372–381. Washington, D.C.: National Tax Association, 1996.

Burtless, Gary.
"Introduction and Summary." In *Does Money Matter?*, edited by Gary Burtless, 1–23. Washington D.C.: The Brookings Institution, 1996.

Card, David, and Alan Krueger.
"Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100 No. 1 (February, 1992): 1–40.

Card, David, and Abigail Payne.
"School Finance Reform, the Distribution of School Spending and the Distribution of

SAT Scores." NBER Working Paper No. 6766. Cambridge, MA: National Bureau of Economic Research, 1998.

Carnoy, Martin.
School Vouchers: Examining the Evidence. Washington, D.C.: Economic Policy Institute, 2001.

Carnoy, Martin, and Susanna Loeb.
"Does External Accountability Affect Student Outcomes? A Cross–State Analysis." *Education Evaluation and Policy Analysis* 25 No. 1 (Winter, 2003).

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor.
"Segregation and Resegregation in North Carolina's Public School Classrooms." Working Paper Series SAN 02-03, August, 2002.

Collins, Ann, Jean Layzer, J. Lee Kreader, Alan Werner, and Fred Glantz.
National Study of Child Care for Low–Income Families: State and Community Substudy Interim Report. Cambridge, MA: Abt. Associates, 2000.

Cullen, Julie Berry.
"The Impact of Fiscal Incentives on Student Disability Rates." *Journal of Public Economics* (forthcoming).

Cullen, Julie Berry, and Susanna Loeb.
"School Finance Reform in Michigan: Evaluating Proposal A." In *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity*, edited by William Duncombe and John Yinger. Cambridge, MA: MIT Press, forthcoming.

Cullen, Julie, Brian Jacob, and Steven Levitt.
"The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools." NBER Working Paper No. 7888. Cambridge, MA: National Bureau of Economic Research, 2000.

Cullen, Julie, Brian Jacob, and Steven Levitt.
""The Impact of School Choice on Enrollment and Achievement: Evidence from over 1,000 Lotteries." University of Michigan Working Paper. Ann Arbor, MI: University of Michigan, 2002.

Cullen, Julie, and Randall Reback.
"Tinkering Towards Accolades: School Gaming under a Performance Accountability System." University of Michigan Working Paper. Ann Arbor, MI: University of Michigan, 2002.

Dee, Thomas S.
"The Capitalization of Education Finance Reforms." *Journal of Law and Economics* 43 No. 1 (April, 2000): 185–214.

Dee, Thomas S.
"Standards and Student Outcomes: Lessons from the 'First Wave' of Education Reform." Swarthmore College Working Paper. Swarthmore, PA: Swarthmore College, 2002.

Downes, Thomas A.
"Evaluating the Impact of School Finance Reform on the Provision of Public Education: The California Case." *National Tax Journal* 45 (1992): 405–19.

Downes, Thomas A.
"School Finance Reform and School Quality: Lessons from Vermont." In *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity*, edited by William Duncombe and John Yinger. Cambridge, MA: MIT Press, forthcoming.

Downes, Thomas A., and David N. Figlio.
"Do Tax and Expenditure Limits Provide a Free Lunch? Evidence on the Link Between Limits and Public Sector Service Quality." *National Tax Journal* 52 No. 1 (March, 1999): 113–28.

Downes, Thomas A., and David Schoeman.
"School Finance Reform and Private School Enrollment Evidence From California." *Journal of Urban Economics* 43 No. 3 (May, 1998): 418–43.

Duncombe, William, and Jocelyn Johnston.
"Is Something Better Than Nothing: An Assessment of School Finance Reform in Kansas." In *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity*, edited by William Duncombe and John Yinger. Cambridge, MA: MIT Press, forthcoming.

Duncombe, William, and John Yinger.
"Performance Standards and Educational Cost Indexes: You Can't Have One Without the Other." In *Equity and Adequacy in Education Finance: Issues and Perspectives*, edited by Helen F. Ladd, Rosemary Chalk, and Janet S. Hansen, 260–97. Washington, D.C.: National Academy Press, 1999.

Eberts, Randall W., and Kevin M. Hollenbeck. *An Examination of Student Achievement in Michigan Charter Schools*. Upjohn Institute Staff Working Paper No. 01-68. Kalamazoo, MI: Upjohn Institute, 2001.

Ferguson, Ronald.
"Paying for Public Education: New Evidence on How and Why Money Matters." *Harvard Journal on Legislation* 28 (Summer, 1991): 465–97.

Figlio, David N., and Lawrence S. Getzler.
"Accountability, Ability and Disability: Gaming the System." NBER Working Paper No. 9307. Cambridge, MA: National Bureau of Economic Research, 2002.

Figlio, David, and Jens Ludwig.
"Sex, Drugs and Catholic Schools: Private Schooling and Non–Market Adolescent Behaviors." NBER Working Paper No. 7990. Cambridge, MA: National Bureau of Economic Research, 2000.

Flanagan, Ann, and Sheila Murray.
"A Decade of Reform: The Impact of School Reform in Kentucky." In *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity*, edited by William Duncombe and John Yinger. Cambridge, MA: MIT Press, forthcoming.

Frederiksen, Norman.
The Influence of Minimum Competency Tests on Teaching and Learning. Princeton, NJ: ETS Policy Information Center, 1994.

Gill, Brian P., P. Michael Timpane, Karen E.
Ross, and Dominic J. Brewer.
Rhetoric Versus Reality: What We Know and What We Need to Know About Vouchers and Charter Schools. Santa Monica, CA: RAND Corporation, 2001.

Goldhaber, D., D. Perry, and E. Anthony.
"NBPTS Certification: Who Applies and What Factors are Associated with Success?" University of Washington and Urban Institute Working paper, 2002.

Greene , Jay P., Paul E. Peterson, and Jiangtao Du.
"Effectiveness of School Choice: The Milwaukee Experiment." Program in Education Policy and Governance, Harvard University Occasional Paper 97–1. Cambridge, MA: Harvard University, 1997.

Grogger, Jeff.
"School Expenditures and Post–Schooling Earnings: Evidence from High School and Beyond." *The Review of Economics and Statistics* 78 No. 4 (November, 1996): 628–37.

Gronberg, Timothy, and Dennis W. Jansen.
"Navigating Newly Chartered Waters: An Analysis of Texas Charter School Performance." Report from the Texas Public Policy Institute. Texas A&M University, Texas Public Policy Foundation, 2001.

Guilfoyle, Jeffrey R.
"The Effect of Property Taxes and School Spending on House Prices: Evidence from Michigan's Proposal A." Michigan Department of Treasury. Mimeo, 1998.

Hanushek, Eric.
"Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects." University of Rochester and NBER Working paper, 1999.

Hanushek, Eric, John Kain, and Seven Rivkin.
"Teachers, Schools and Academic Achievement." NBER Working Paper No. w3691. Cambridge, MA: National Bureau of Economic Research, August 1998.

Hanushek, Eric A., John F. Kain, and Steven
G. Rivkin.
"The Impact of Charter Schools on Academic Achievement." Working paper, 2002.

Hanushek, Eric A., and Steven G. Rivkin.
"Does Public School Competition Affect Teacher Quality?" In *The Economics of School Choice*, edited by Caroline Minter Hoxby. Chicago: University of Chicago Press, 2003.

Hanushek, Eric A., Steven G. Rivkin, and Lori
L. Taylor.
"Aggregation and the Estimated Effects of School Resources." *The Review of Economics and Statistics* 78 No. 4 (November, 1996): 611–27.

Heckman, James, Anne Layne–Farrar, and
Petra Todd.
"Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings." *Review of Economics and Statistics* 78 No. 4 (November, 1996): 562–610.

Howell, William F., Patrick J. Wolf, Paul E. Peterson, and David E. Cambell.
"Test–Score Effects of Vouchers in Dayton, Ohio, New York City, and Washington D.C." Evidence from Randomized Field Experiments." Program in Education Policy and Governance, Harvard University, Working Paper. Cambridge, MA: Harvard University, 2000.

Hoxby, Caroline M.
"How Teachers' Unions Affect Education Production." *The Quarterly Journal of Economics* 111 No. 3 (August, 1996): 671–718.

Hoxby, Caroline M.
"The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *The Quarterly Journal of Economics* 115 No. 4 (November, 2000a): 1239–85.

Hoxby, Caroline M.
"Does Competition Among Public Schools Benefit Students and Taxpayers?" *The American Economic Review* 90 No. 5 (December, 2000b): 1209–38.

Hoxby, Caroline M.
"All School Finance Equalizations are Not Created Equal." *The Quarterly Journal of Economics* 116 No. 4 (November, 2001): 1189–231.

Hoxby, Caroline M.
"Would School Choice Change the Teaching Profession?" *Journal of Human Resources* 37 No. 4 (Fall, 2002): 846–91.

Imazeki, Jennifer, and Andrew Reschovsky.
"School Finance Reform in Texas: A Never Ending Story?" In *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity*, edited by William Duncombe and John Yinger. Cambridge, MA: MIT Press, forthcoming.

Jacob, Brian A.
"Getting Tough? The Impact of Mandatory High School Graduation Exams on Student Achievement and Dropout Rates." *Education Evaluation and Policy Analysis* 23 No. 2 (Summer, 2001): 99–122.

Jacob, Brian A.
"Accountability, Incentives, and Behavior: The Impact of High–Stakes Testing in the Chicago Public Schools." NBER Working Paper No. 8968. Cambridge, MA: National Bureau of Economic Research, Fall 2002.

Jacob, Brian A., and Lars Lefgren.
"Remedial Education and Student Achievement: A Regression–Discontinuity Analysis." NBER Working Paper No. 8918. Cambridge, MA: National Bureau of Economic Research, 2002.

Jacob, Brian A., and Steven D. Levitt.
"Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." NBER Working Paper No. 9413. Cambridge, MA: National Bureau of Economic Research, 2003.

Jespsen, Chris, and Steven G. Rivkin.
Class Size Reduction, Teacher Quality, and Academic Achievement in California Public Elementary Schools. San Francisco, CA: Public Policy Institute of California, 2002.

Krueger, Alan B.
"Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 No. 2 (May, 1999): 497–532.

Krueger, Alan B., and Diane M. Whitmore.
"The Effect of Attending a Small Class in the Early Grades on College–Test Taking and Middle School Test Results: Evidence from Project STAR." NBER Working Paper #427. Cambridge, MA: National Bureau of Economic Research, 1999.

Krueger, Alan B., and Pei Zhu.
"Another Look at the New York City Voucher Experiment." NBER Working Paper No. w9418. Cambridge, MA: National Bureau of Economic Research, January 2003.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff.
"Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis." *Education Evaluation and Policy Analysis* 24 No. 1 (Spring, 2002): 37–62.

Loeb, Susanna.
"Estimating the Effects of School Finance Reform: A Framework for a Federalist System." *Journal of Public Economics* 80 No. 2 (May, 2001): 225–47.

Loeb, Susanna, and John Bound.
"The Effect of Measured School Inputs on Academic Achievement: Evidence from the

1920s, 1930s and 1940s Birth Cohorts." *Review of Economics and Statistics* 78 No. 4 (November, 1996): 653–64.

Loeb, Susanna, and Marianne Page.
"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 No. 3 (August, 2000): 393–408.

Metcalf, Kim.
Evaluation of the Cleveland Scholarship and Tutoring Grant Program: 1996–1999. The Indiana Center for Evaluation, Indiana University, 1999.

Murray, Sheila E., William N. Evans, and Robert M. Schwab.
"Education–Finance Reform and the Distribution of Education Resources." *The American Economic Review* 88 No. 4 (September, 1998): 789–812.

Neal, Derek.
"The Effects of Catholic Secondary Schools on Educational Achievement." *Journal of Labor Economics* 15 No. 1 (January, 1997): 98–123.

Peterson, Paul E., William G. Howell, Patrick J. Wolf, and David E. Campbell.
"School Vouchers: Results from Randomized Experiments." In *The Economics of School Choice*, edited by Caroline Hoxby. Chicago, IL: University of Chicago Press, 2003.

Reback, Randall.
"Capitalization under School Choice Programs: Are the Winners Really the Losers?" National Center for the Study of Privatization in Education Occasional Paper #66. Ann Arbor, MI: University of Michigan, 2002.

Rose, Heather, and Julian R. Betts.
Math Matters: The Links Between High School Curriculum, College Graduation and Earnings. San Francisco, CA: Public Policy Institute of California, 2001.

Rouse, Cecilia Elena.
"Schools and Student Achievement: More Evidence from the Milwaukee Choice Program." *Quarterly Journal of Economics* 113 No. 2 (1998): 553–602.

Schwartz, Amy Ellen, Leanna Stiefel, and Dae Yeop Y. Kim.
"The Impact of School Reform on Student Performance: Evidence from the New York Network for School Renewal." *Journal of Human Resources* (forthcoming).

Solmon, Lewis, Kern Paark, and David Garcia.
"Does Charter School Attendance Improve Test Scores? The Arizona Results." The Goldwater Institute, University of Arizona, 2001.

Sonstelie, Jon, Eric Brunner, and Kenneth Ardon.
For Better or For Worse? School Finance Reform in California. Public Policy Institute of California. 2000.

Stinebrickner, Todd.
"An Analysis of Occupational Change and Departures from the Labor Force: Evidence of the Reasons Teachers Quit." *Journal of Human Resources* 37 No. 1 (Winter, 2002): 192–216.

Wexler, Edward, Jo Ann Izu, Lisa Carlos, Bruce Fuller, Gerald Hayward, and Michael Kirst.
California's Class Size Reduction: Implications for Equity, Practice, and Implementation. San Francisco, CA: WestEd, 1998.

Witte, John F., Troy D. Sterr, and Christopher A. Thorn.
"Fifth–Year Report: Milwaukee Parental Choice Program." University of Wisconsin. Mimeo, 1995.

Wolf, Patrick J., Paul E. Peterson, and Martin R. West.
"Results of a School Voucher Experiment: The Case of Washington, D.C. After Two Years." Program on Education Policy and Governance, Harvard University Working Paper. Cambridge, MA: Harvard University, 2001.