Chapter 7: Educational Reforms
Susanna Loeb and Patrick J. McEwan

1. Introduction

Over 55 million children and youth attend elementary and secondary schools in the United States, 89 percent in public schools. These students spend approximately 1,000 hours each year in schools across the country, for which local, state and federal governments spent over $550 billion (NCES, 2008).[1] Education is an intensive and costly enterprise. It also has the potential to dramatically improve the opportunities for students. In the United States, estimates of the return to an additional year of schooling are in the neighborhood of 10 percent, depending on the data and method (Card, 1999).[2] Educational attainment is also associated with differences in individual health, incarceration, and dependence on public assistance (Belfield and Levin, 2007).

While schooling improves children's lifetime opportunities, the evidence on how to use scarce time and resources to maximize children's outcomes while in school is less clear. Higher spending, lower class sizes, and additional computers can have little effect on student learning if not implemented well. As an example, the early, non-experimental literature was famously interpreted as providing little evidence that "money mattered" in raising test scores (Hanushek, 1986, 2006). The summarized studies in these reviews were often methodologically weak, and other reviewers turned up conflicting interpretations of the evidence (Greenwald, Hedges, and

[1] See http://nces.ed.gov/pubs98/yi/y9638a.asp for hours information.
[2] Economists have long worried that estimates of the return to schooling do not have a causal interpretation. High-ability individuals may earn more, in addition to being more likely to continue in school, perhaps leading to a spurious association between schooling and wages. A large literature, including twins studies and other attempts to isolate exogenous variation in schooling, rarely suggest that the return to years of schooling is biased upward. Indeed, they frequently yield even larger estimated returns, perhaps because the methods estimate returns for a unique subpopulation (Card, 1999).

Laine, 1996; Krueger, 2003). Still, the debate helped shift researchers' attention away from spending levels to how the dollars are spent. In particular, economists increasingly focused their work on policies designed to improve schools' incentives to use resources efficiently, whether by enhancing parental school choice (via private school vouchers or charter schools) or by test-based accountability rules.

In the last decade, the quality of education research has improved with the development of new methods and the increased availability of relevant data (Angrist, 2004; Barrow and Rouse, 2005). In the late 1990s, economists published influential re-analyses of experimental data on the impact of class size reduction in Tennessee (Krueger, 1999) and private school vouchers in Milwaukee (Rouse, 1998). The next 10 years brought an avalanche of new and better research, catalyzed by three factors. First, the available data have improved, especially with collection of longitudinal administrative data on students in several U.S. states and cities (Loeb and Strunk, 2003). Second, formerly "hypothetical" policies—especially related to choice and accountability—have actually been implemented and studied (Figlio and Ladd, 2008; Zimmer and Bettinger, 2008). Third, the U.S. Department of Education and some foundations increasingly require the use of research designs able to yield credible causal findings, especially randomized experiments and discontinuity designs.

Education polices comprise a vast array of programs and approaches. Many of the policies can be categorized into one of three groups: (1) direct investments in schools, including school improvement grants and class size reductions; (2) interventions that target the teacher workforce through wages, recruitment, or professional development programs; and (3) interventions that aim to increase accountability and change decision-making in schools through either enhancing parental choice or increasing test-based accountability. This chapter selectively

reviews the evidence on the effects of different approaches within each of these three groups, focusing on high-quality studies.

2. Estimating Policy Effects

Economists have traditionally used non-experimental data to estimate education production functions, in which student test scores are regressed on a "kitchen sink" of explanatory variables. These include attributes of students and their families (e.g., ability and income), attributes of teachers and schools (e.g. pre-service training and expenditures), and attributes of peers and communities. The usual goal is to isolate the causal effect of school inputs that can potentially be manipulated by school authorities. The empirical task is complicated by the fact that observed test scores are the cumulative result of investments by families and schools throughout a child's life (Todd and Wolpin, 2003). Only a fraction of these investments are observed in most datasets. It is common, in such cases, to include a lagged test score in regressions as an implicit control for prior family or school variables. Even if this does control for prior influences on test scores, the models must fully control for contemporaneous factors associated with student participation in different policies or programs, and this is very difficult. Researchers are often left wondering whether their regressions effectively adjust for the selection of different students into different policy environments.

Alternatively, researchers attempt to identify "clean" variation in policy variables (like class size) that is uncorrelated with unobserved variables that affect test scores. In experiments, the researcher randomly assigns a subset of students, classrooms, or schools to receive a policy treatment, and randomly denies it to others. By design, in large studies this randomization ensures that treated subjects are similar to untreated ones, except for their exposure to the policy, and that subsequent comparisons of outcomes will likely yield unbiased effects. In a few cases,

randomized assignment is a natural byproduct of program implementation, as in lotteries to allocate private school vouchers.

When randomization is not feasible, it is sometimes possible to identify variation in policy treatments that is "as good as random." Among the many varieties of quasi-experiments, the regression-discontinuity design can yield convincing causal results (Shadish, Cook, and Campbell, 2002; Angrist and Krueger, 1999). Treatments are assigned on the basis of a single variable and an assignment cutoff (i.e., schools receive a program if their poverty rate is below a fixed threshold, but not above). Assuming that schools or students on either side of the cutoff are otherwise similar, comparisons of the two groups' outcomes are a reasonable estimate of the causal effect. It is akin to very local randomized experiment (Lee, 2008).

In the next sections, we review research that mainly uses experimental and discontinuity research designs. Still, it bears emphasis that our goal is to generalize these effects beyond the immediate research setting, and that doing so is sometimes more art than science. First, policy effects may be heterogeneous across students. If effects are heterogeneous, then randomized experiments succeed in identifying the average effect among students (or occasionally within subgroups of students in large experiments). However, the research participants are often "special" in ways that could increase or decrease their treatment effects, relative to the typical student that the real-world policy would eventually target. Experimental subjects often volunteer to be randomly assigned, are drawn disproportionately from a particular race or income-level, or reside in compact geographic areas with unique institutions (e.g. school finance and accountability rules). Regression-discontinuity studies face a stricter version of this problem because they can only identify local average effects for the subpopulation or students or schools in the vicinity of the assignment cutoff. Often this is policy-relevant, since decision-makers

might raise or lower eligibility cutoffs. But for broader decisions about the cost-effective targeting of resources, it would be useful to understand whether treatment effects are different for subjects far away from eligibility cutoffs (i.e., the poorest schools that qualify for Title I funds, rather than "just poor enough" schools).

Second, the best causal research is frequently conducted on a small scale. However, scaling up an intervention can provoke unanticipated general equilibrium effects. Sometimes this severely undermines a policy's original objectives. The best-known case in education is California's statewide class size reduction in the late 1990s (Jepsen and Rivkin, 2002), itself directly inspired by Tennessee's small-scale experiment (Krueger, 1999). In other cases, the potential general equilibrium effects in scale-ups are of greater policy interest than the treatment effects actually identified in the small-scale research. For example, private school voucher experiments identify the effects of private school attendance on the few students who are offered vouchers. Yet, most policy-makers are at least as interested in how a large voucher offer (and the concomitant re-shuffling of students across schools) would affect the outcomes of all students through increasing market competition or school stratification (Hsieh and Urquiola, 2006; Hoxby, 2000b). We keep these potential limitations of small scale causal studies in mind in the review that follows.

3. Direct Resource Investments

For many years, the debate on the effects of resource investments in schools was dominated by Hanushek's (1986, 2006) summaries of non-experimental research findings which identified no evidence that either per-pupil expenditures or pupil-teacher ratios (a class size proxy) were systematically associated with student achievement. In the ensuing debate, Greenwald, Hedges, and Laine (1996) argued, on the contrary, that the same evidence revealed

consistent effects of resources.[3] The median coefficient from the collection of studies implies that $500 per student (1994 dollars) is associated with a 0.15 standard deviation increase in performance. Krueger (2003) re-weights studies and finds that the pupil-teacher ratios do appear to affect achievement.

This early debate has been rendered moot by the increasing availability of higher-quality studies that are more tightly focused on relevant policy alternatives. As Krueger (1999, p. 528) notes, "one well-designed experiment should trump a phalanx of poorly controlled, imprecise observational studies based on uncertain statistical specifications." The next sections review the best recent evidence on four types of resource investments. First, we consider three policies that affect the level of per-pupil revenues or expenditures in schools: the Federal Title I program which directs additional funds to high-poverty schools; a California policy of school bonuses for high-performing schools; and a range of school equity reforms that leveled up (or down) the expenditures in schools. Second, we briefly review the best evidence on whether class size reduction increases test scores. Third, we review whether specialized instructional packages can raise test scores, focusing on the Success for All reading program. Fourth, we consider whether computer-assisted instruction causes test score improvements.

A. Dollars

Title I is the largest Federal education program in K-12 education, with $12 billion allocated in fiscal year 2005 (van der Klaauw, 2008). Besides its scale, it is notable for its objective targeting of resources towards counties and schools with larger numbers of poor students. Title I's distribution rule is intended to promote a transparent and well-targeted

[3] Combined significance tests reject the null hypothesis of no positive effects, and accept the null of no negative effects.

resource allocation, but it also facilitates a regression-discontinuity design.[4] van der Klaauw (2008) applies this strategy to school-level data on New York City public schools. Schools with poverty rates below a threshold have sharply higher probabilities of receiving Title I funds (about five percent of a school's budget), but are otherwise similar to schools just above the threshold. The author finds that Title I designation did not produce achievement gains in 1993, 1997, or 2001, and may even have led to achievement declines in the earlier years. However, Title I schools also do not appear to have higher expenditures, perhaps because the State or local authorities remove other funds (for related evidence, see Gordon, 2004). van der Klaauw further notes that a popular use for Title I funds was "pull-out" remedial instruction. Despite its easier compliance with federal rules, it has little demonstrated effectiveness as an instructional strategy.

In 2000 and 2001, California offered financial rewards to schools that met specified achievement targets (Bacolod et al., 2008). Upon winning, schools received one-time, unrestricted bonuses that amounted to about five percent of per-pupil expenditures. Though apparently intended for computers or other instructional purposes, it appears that most funds were returned to teachers in the form of bonuses. Using a discontinuity approach, Bacolod et al. (2008) compare subsequent achievement of schools that barely qualify for an award with those that barely miss one. They find no gains in student achievement.

Finally, we consider school finance reforms, which constitute one of most significant attempts in the last 30 years to influence the resources available to schools enrolling

[4] Similar evaluation strategies have been applied to programs in Chile and the Netherlands that allocated additional resources to schools based on measures of achievement or disadvantage (Chay, McEwan, and Urquiola, 2005; Leuven et al., 2007). The Chilean program found moderately positive test score effects of intensive after-school tutoring, while the Dutch program found some negative effects of extra funding for computers and software.

disadvantaged children.[5] Most reforms were mandated by state courts, following successful challenges to the state constitutionality of locally-based systems of school finance. Because these "experiments" were initiated by courts and legislatures, and not researchers, their causal effects are harder to identify. Corcoran and Evans (2008) compare the evolution of expenditures in states with and without reforms, finding that finance reforms typically reduced within-state inequality in per-pupil expenditures by 15 to 19 percent. Further, this does not appear to have occurred through a simple "leveling down" of higher-spending schools.

Although one anticipates that additional resources should affect student outcomes, there is mixed evidence on this fundamental question (Corcoran and Evans, 2008). In a cross-state analysis, Card and Payne (2002) find that states with court-mandated reforms experienced reductions in test score inequality, but the researchers are hampered by the use of SAT scores that are taken by a subset of students. Other authors working with cross-state data find no effects (Downes and Figlio, 1998).

The most credible studies are typically conducted in a single or small number of states, but here too the evidence is conflicting (Corcoran and Evans, 1998). Researchers have found no effect on test scores in Kentucky (Flanagan and Murray, 2004), but positive effects on pass rates in Michigan (Papke, 2005; Cullen and Loeb, 2004). Of state-specific studies, Guryan's (2003) is one of the most convincing. His discontinuity study assesses how increased spending, induced by changes in the Massachusetts school funding formula, affected test scores. Specifically, he relies on spending variation created when districts fall on one side or another of funding thresholds established by the formula. He finds that increasing per-pupil expenditures by $500

[5] Another class of finance reform, not considered here, is tax limitations, which removed resources from schools (Downes and Figlio, 2008).

per student[6] (about half a standard deviation) yields tests score increases in 4th grade mathematics and reading of roughly 0.06-0.15 standard deviations, respectively.[7] The eighth grade test score results are also positive but not robust to alternative specifications.

B. Class Size Reduction

Given the popularity of class size reduction, there are surprisingly few high-quality studies of its effects on test scores. As a result, most attention has focused on a large randomized experiment conducted in Tennessee during the mid-1980s (Schanzenbach, 2007).[8] Within 79 volunteer schools, researchers randomly assigned students and teachers to "small" classes (13-17 students) and "regular" classes (22-25 students).[9] This treatment was maintained for four years (between Kindergarten and third grade), though not all students participated in all years. For example, some new students entered the school after Kindergarten, and a small proportion moved between classes within schools.

On average, the Tennessee results suggest that students who were initially assigned to smaller classes have test score gains of 0.15 standard deviations by the end of third grade, though similar achievement gains were already in evidence by the end of Kindergarten. The effects are even larger for the subset of black students, or lower-income students. In follow-up studies, these effects were much smaller and statistically insignificant by the end of eighth grade. In a

[6] These appear to be 1996 dollars.
[7] The coefficients estimates are from the fixed-effects specification in Table 4, column 2. Guryan (2003) divides these coefficients by the standard deviation of district-level test score means, which tends to inflate effect sizes when between-district test score variation is small. To make the effect size comparable to others, and in the absence of a student-level standard deviation of tests score, we assume it to be (district-level SD)/sqrt(intra-district correlation coefficient), where the second term in parentheses is the proportion of variance in test scores accounted for by between- rather than within-district variation (What Works Clearinghouse, 2007). We assume it to be 0.2, which is likely over-stated.
[8] For earlier analyses of the Tennessee experiment, see Krueger (1999) and Krueger and Whitmore (2001).
[9] They also considered a third group, consisting of regular classes with teachers' aides.

surprising finding, however, it appears that students eventually had a higher probability of taking a college entrance exam (0.02), again larger for black or lower-income students.

Despite these findings, reducing class size can be a costly endeavor. Following Schanzenbach (2007), we can assume that a seven student reduction in the typical class size increased per-pupil expenditures ($10,551 in 2005) by 47 percent, an annual per-pupil increase of $4,959. The intervention lasted four years, but the average student participated for 2.3. Assuming a three percent discount rate and inflating dollar estimates to 2007, the discounted per-pupil cost of the STAR intervention is $11,865. This is just under $16,000 per 0.2 standard deviation gain in test scores (but just under $10,000 per 0.2 among black students).

In another study of class size, Hoxby (2000a) finds no class size effects in Connecticut, using different quasi-experimental approaches, including variation in class size driven by plausibly random changes in the size of local populations. She further implements a discontinuity analysis, using sharp decreases in class size caused when enrollments exceed specified caps. The evaluation approach has been applied in several other countries, notably Israel (Angrist and Lavy, 1999) and Bolivia (Urquiola, 2006), showing positive tests score effects of reducing class size. However, Urquiola and Verhoogen (forthcoming) suggest that discrete enrollments can be manipulated by schools, perhaps violating identifying assumptions of the discontinuity design in these settings.

C. Curriculum and Instructional Programs

To many educators and parents, an obvious avenue for improving schools is to improve the curriculum and instruction offered by schools. There are hundreds of different curricular and instructional reform approaches. Many are piecemeal “add-ons” to existing school programs,

few of which are supported by high-quality studies. Some of these approaches, however, are "whole-school" reforms that consist of comprehensive and coordinated efforts to overhaul the curriculum, instruction, technology, training, and other aspects of school operations (Levin, 2002). These reforms are quite varied in their strategies and goals. A few have been subjected to rigorous evaluation, but overall the research literature in this area is sparse.

As an example, randomized experiments have assessed the effectiveness of the School Development Program of James Comer finding mixed results on student achievement (Cook et al., 1999; Cook et al., 2000). However, these experiments were hampered by relatively small samples of participating schools, which were the unit of randomization. A quasi-experimental, interrupted time-series analysis of Henry Levin's Accelerated Schools Project showed positive effects , but the studylacked an untreated comparison group to verify the robustness of these results (Bloom, 2003). In New York City, researchers have compared achievement over time in a varied group of reform schools (without random assignment) to non-reform schools. This research yields mixed achievement results, and it is unclear whether non-reform schools are an adequate comparison group (Bifulco, Duncombe, and Yinger, 2004; Schwartz, Stiefel, and Kim, 2004).

To date, the most rigorous evaluation of curriculum and instructional programs has been conducted on the Success for All reform, which focuses on improving reading skills. Success for All is a package of materials, training, and a scripted blueprint for implementing the program, generally targeted at high-poverty and low-achieving schools (Borman et al., 2007). In a random assignment study, 41 schools were randomly assigned to apply the reform (or not) in early grades. After three years, the reading scores of students in Success for All treatment schools were 0.21 to 0.36 standard deviations higher than students in the control schools, depending on

the test. Borman and Hewes (2002) estimate that Success for All has annual per-pupil costs of $795 (in 2000).[10] Assuming a discount rate of three percent in a three-year intervention, and inflating dollars to 2007, the discounted per-pupil cost of the intervention is $2,789. Thus, depending on the effect size estimate, it costs from $1500-$2600 per 0.2 standard deviations.

D. Computer-Assisted Instruction

Many countries and states have embarked on costly plans to increase the number of computers in schools, ranging from placement of computers in classrooms to thoughtful efforts integrate computers into schools' instructional plans. A small number of high-quality studies have assessed the extent to which the latter efforts have a causal effect student learning. The mixed evidence suggests that results depend vitally on the program details.

As examples, two randomized experiments have tested the effects of the "Fast ForWord" program, a popular computer-based approach to raising reading and language ability (Borman and Benson, Forthcoming; Rouse and Krueger, 2004). Neither finds meaningful effects for the program. A large, federally-funded randomized experiment also finds no effects, using a diverse array of instructional products in both math and reading (Dynarski et al., 2008). This study randomly assigned teachers within 132 schools to use one of 16 of computer-based approaches. After one year, there were no detectable test score effects.

In contrast, a recent evaluation of a computer-based algebra program ("I Can Learn") found encouraging effects on student test scores (Barrow, Markman, and Rouse, 2008). The researchers randomly assigned teachers (or class periods) within schools, roughly following the design of Dynarski et al. (2008), and identified test score effects of 0.17 standard deviations.

[10] Barnett (1996) reports slightly lower per-pupil costs. They may under-estimate full social costs, because Success for All incurs opportunity costs for volunteering parents and existing staff (King, 1994).

Barrow et al. (2008) calculate a per-student intervention cost of $283 for a single variety of computer-assisted instruction in math, or $333 per 0.2 standard deviations.[11]

E. Summary

Research on direct investments in schools finds great variation in effects. Given the vastly improved quality of these studies (relative to even five years ago), the mixed patterns of evidence cannot be attributed entirely to bad methods or data. Rather, it suggests that the debate has (usefully) shifted to questions of how and when resources matter for student outcomes, rather than whether they matter at all.

Most evidence on increases in per-pupil expenditures does not show test score improvements for students; however, this lack of impact may partly reflect funds being used for ineffective interventions such as pull-out tutoring or one-time bonuses. The literature on school finance reforms suggests that the subsequent increases in funding in formally low-spending areas may have diminished test score inequality, but our understanding of how these gains occurred or failed to occur is surprisingly modest. Further progress rests on obtaining a more nuanced understanding of how resources are used in specific policy settings.

Class size reduction can have positive effects on student learning, but at substantial costs. There is no shortage of innovative attempts to reform curriculum and instruction, but few have been rigorously evaluated. Still, it appears that intensive efforts to improve reading skills can successfully raise test scores. Computers also are no panacea for schools especially in the absence of clear instructional goal), but a well-conceived math program that integrates computers can demonstrate robust effects in just one year.

[11] Though the upfront costs of a computer lab and training are relatively high, they are amortized across 7 years.

4. Teachers and Teaching

Schools spend more on teachers than on any other budget category, and there is strong evidence that these expenditures can affect student achievement. Comparing teachers within the same schools, Rivkin, Hanushek, and Kain (2005) find that a one standard deviation increase in average teacher quality for a grade raises average student achievement in the grade by at least 0.11 standard deviations of the total test score distribution in mathematics and 0.095 standard deviations in reading. Yet, knowing that teachers vary meaningfully in their effectiveness does not provide a roadmap for how to increase teacher quality. In this section we summarize the current knowledge of the effects of three types of policies aimed at improving teacher quality: wage increases, recruitment, and professional development.

A. Wages

Teachers' choices about jobs are responsive to wages. A large literature finds that teachers are more likely to choose teaching when starting wages are high relative to wages in other occupations. Approximately 16.5 percent of public school teachers who decided to move to another school between 2003-04 and 2004-05 reported having done so for better salary or benefits (NCES Schools and Staffing Surveys). For those who left teaching in 2004-05, nearly 15 percent cited salary related reasons. Teacher wages have increased dramatically over the last forty years. Nevertheless, since the 1970s, they have fallen behind salaries in non-teaching jobs for individuals with similar qualifications. Lawyers, doctors, scientists, and engineers earn substantially more, as do managers and sales and financial service workers (Corcoran, Schwab and Evans, 2004). Bacolod (2007) finds that highly-qualified teachers are especially sensitive to changes in relative wages. The less teachers are paid, relative to professionals, the less likely high-abilitywomen are to choose teaching. The opportunity cost of becoming a teacher, in terms

of salary forgone in alternative professions, is high. However, teachers likely work fewer hours and fewer days, at least partially compensating for this forgone income. In 2003-04, the average base salary of regular full-time teachers was $44,400 per annum.

While the evidence on the effects of wages on teachers' decisions is persuasive, high-quality evidence on the effects of teacher wage increases on students is sparse. Loeb and Page (2000) use state-level panel data from the 1960-1990 Public Use Microdata Samples from the U.S. Census to examine changes in teacher wages over time. They identify the effect of wages from both changes in relative teacher salaries and changes in only the salaries of non-teaching college graduates, the opportunity cost of becoming a teacher. The study finds that increases in teacher wages of 10 percent led to a 3 to 4 percent drop in student dropout rates and a 1 to 2 percent increase in college enrollment. The authors' simple calculations suggest that the benefits of a 10 percent wage increase would slightly outweigh the costs.

The Loeb and Page study examines the effects of average wage increases, but wage increases can also be targeted to specific needs and outcome goals. Conceptually, directly linking wage increases to improved outcomes for students is a logical means of maximizing their effects. By paying teachers more when their students learn more, performance-based pay creates incentives for teachers to focus their efforts on student learning and it can create incentives for the most effective teachers to enter or remain in the teaching professions. There are also potential drawbacks of performance-based pay. We do not measure all aspects of student learning that we care about; and, thus, by creating incentives to focus on the measured outcomes we may be hurting students on unmeasured dimensions. Similarly, it is difficult to create performance-based pay systems that provide teachers with incentives to treat their students equitably. The reward formulas often make it beneficial to concentrate more on some students,

perhaps those who are performing quite close to an achievement cutoff, to the detriment of other students. In addition, if cooperation among teachers is important to student learning, then performance-based pay systems can have detrimental effects if they reduce incentives for teachers to cooperate.

There is very little solid evidence on performance-based pay in the United States, so we briefly discuss the higher-quality and mixed evidence from developing countries. Two studies use experimental methods to estimate the effects of performance pay for teachers in India and Kenya. Muralidharan and Sundararaman (2006) report effects from a randomized experiment in 500 schools in the rural Indian state of Andhra Pradesh. The schools were divided into five groups: the control group, schools with individual teacher bonuses tied to student test-score gains, school-based bonuses, teacher aides, and extra funds. The average bonus was approximately four percent of average salary but could reach a maximum of 29 percent for the individual bonuses and 14 percent for the school-based bonuses. The study finds that students in schools with either incentive program performed better than those in the other schools. Relative to the control schools, these students gained 0.19 and 0.12 standard deviations more in math and language tests respectively.

Glewwe, Ilias, and Kremer (2003) implement a smaller experiment in 100 rural schools in Kenya. In this case, all the bonuses were school-wide and represented approximately 21 to 43 percent of teacher wages. The authors found that students in schools with merit bonuses were more likely to pass their exams during the two years of the program but that the students did not perform better in subsequent years. In addition, the researchers found little evidence that teachers increased their effort or focus on instruction as a result of the program. It is clearly difficult to generalize from rural India and Kenya to schools in the United States. Current

performance-based pay programs in Denver, Nashville and other cities are likely to provide useful evidence on this approach in the relatively near future.

B. Recruitment

Wage changes are a straightforward means of affecting the teacher workforce but they are not the only means and they may not be the most cost-effective. Teach for America (TFA) and other recruitment programs such as the New York City Teaching Fellows have demonstrated that recruitment combined with reorganization of the timing of entry requirements for teaching can drastically change the pool of teacher candidates. As an example, for the 2006 school year, TFA received approximately 19,000 applications for approximately 2,400 openings received, including 10 percent of the senior classes at Spelman and Yale and 8 percent of the senior class from the California Institute of Technology (Teach for America, 2006).

Studies of the effects of Teach for America teachers on student achievement have tended to find more positive effects in math than in reading or English language arts, and more positive effects when comparing TFA teachers to the average teacher in the school than to teachers who obtained certification through traditional teacher education programs. Decker, Mayer, and Glazerman (2004) designed a within-school random assignment study in 17 schools (100 classrooms) during the 2002-2003 school year. They found that the test scores of students of TFA teachers improved by approximately 0.15 standard deviations more in math than other those of other students in the school. They found no difference in reading.

Teach for America teachers are paid by the district in which they work, as are other teachers. However, there are additional program costs. TFA reports that it must raise $20 million annually to support 1,000 members in New York City schools (some of which may be

reimbursed by school districts).[12] Of these funds, 21 percent goes to recruitment and selection, 21 percent to pre-service training, and 27 percent to professional development. TFA is also a member of AmeriCorps which provides their members with loan forbearance and interest payment on qualified student loans for the two years of participation, and an education award of $4,725 at the end of each year for future educational expenses or to repay qualified student loans. Assuming a typical TFA class size of 18 (Decker et al., 2004), the annual per-pupil cost of supporting a TFA teacher is $1,374 (including TFA's costs and the AmeriCorps stipend). This is roughly $1800 per 0.2 standard deviation in math scores, although it bears emphasis that there are no measured reading effects and these results come from comparing TFA teachers to a range of teachers, many of whom had very little pre-service training.

The Decker study has strong internal validity because students were randomly assigned to teachers within their school. However, because of the variety of non-TFA teachers in schools in the sample the results can be used to draw some conclusions, but not others. For example, it is clear that TFA teachers perform approximately as well in reading and better in math than the other teachers in the school in which they teach, but not necessarily better than teachers who had fulfilled the traditional requirements for teaching. The effects of TFA teachers also may differ across schools and across grade levels, and, thus, the results for elementary schools in the Decker study may not reflect the effects in other contexts.

Several studies have used state and district longitudinal data on students to assess whether TFA teachers produce greater test-score gains among their students than other teachers: two studies in a Texas district, one in rural North Carolina, and two in New York City. These studies confirm some of the Decker study's findings and shed further light on the relative

[12] See https://www.teachforamerica.org/about/regions/new_york_city.htm#financial_sustainability

effectiveness of TFA teachers. Raymond, Fletcher, and Luque (2001) and Darling-Hammond, Holtzman, Gatlin, and Heilig (2005) use data on elementary schools in the same district in Texas and find some positive effects in math but not in reading. Xu, Hannaway and Taylor (2008) is the only study of TFA teachers to assess effects in high school. The authors find that rural North Carolina students of TFA teachers learn more during the course of the year than students of teachers from traditional routes. They estimate that the difference in effectiveness between the routes is approximately equal to twice the difference between the average first-year and average second-year teachers.[13]

Boyd et al. (2006) and Kane, Rockoff and Staiger (2007) study TFA in New York City. Boyd compares TFA teachers to teachers who had completed a traditional teacher certification program. They find that students of TFA teachers gained 0.31 standard deviations *less* in English language arts and about the same in math as traditionally certified teachers in the same schools, though students of TFA teachers did have greater learning gains in math than other not-traditionally-certified teachers such as those who entered teaching through individual evaluation, emergency certified, and other alternative routes.

Teach for America teachers largely replace other not-traditionally-prepared teachers, so the comparison with traditionally prepared teachers may not be the most policy-relevant comparison. As an example, in the Decker et al. experimental study, while only four percent of TFA teachers reported having spent ten or more weeks student teaching compared with 31 percent of other teachers with three or fewer years of experience, all TFA teachers had at least four weeks of student teaching experience during their summer institute, while over half of other

[13] There is not a large enough sample size of TFA teachers in North Carolina high schools to separate the effects by subject area and the results are an average of teachers in algebra I, algebra II, biology, chemistry, geometry, physics, physical science, and English I.

novice teachers had no student teaching experience. Boyd et al. (forthcoming) found that as a result of eliminating emergency certification and implementing intensive recruitment efforts through the New York City Teaching Fellows program and, to a lesser extent, through TFA, the gap between the qualifications of teachers in high-poverty schools and low-poverty schools narrowed substantially between 2000 and 2005. The authors estimate that this change in measured qualifications of teachers alone is likely to have improved the test-score performance of students in the poorest schools approximately 0.03 standard deviations, about half the difference between being taught by a first year teacher and a more experienced teacher.

C. Professional development

Recruitment programs such as TFA concentrate on new teachers, but a variety of professional development policies aim to improve the effectiveness of teachers already in the classroom. The average effect of these policies and programs are not promising. In a summary of this research, Hill (2007, p. 121) writes, "there is little evidence that the system of professional development, taken as a whole, improves teaching and learning in the United States." In one of the best large-scale studies, given its reliance on discontinuity assignment, Jacob and Lefgren (2004) find little evidence that in-service programs in Chicago affected student performance in either math or reading.

There is little argument that professional development programs have not had positive effects on students. Exceptions to this rule seem to appear only when programs are concentrated and intensive. Yoon et al. (2007) reviewed more than 1,300 studies of professional development programs. Of these, only nine met the standards established by the Department of Education's What Works Clearinghouse for estimating causal effects. Combining the results from these studies, the authors conclude that *concentrated* professional development opportunities—in this

case programs that required an average of 49 hours of teacher participation—can improve student achievement by approximately 21 percentile points.

Carpenter et al. (1989) is one example of the studies meeting the criteria in the Yoon report. They randomly assigned 40 first grade teachers to either a control group or a month-long workshop focused on children's development of problem-solving skills in addition and subtraction. Teachers in the control group participated in workshops focused on non-routine problem solving. The program required teachers to attend 20 workshop hours a week for four weeks during the summer and one brief meeting in October, taught by two professors and three graduate students. The researchers found that teacher in who participated in the workshop taught problem solving significantly more and number facts significantly less than did control teachers. Students were given a standardized mathematics achievement pretest in September and a series of posttests in April and May. Students in the treatment group scored approximately 0.4 standard deviations higher on the post-test (the Iowa Test of Basic Skills). This difference, though large, was not statistically significant; however, on the sub-score of complex addition and subtraction, the treatment groups did score a statistically significant 0.5 standard deviations higher.

Using a pre-test/post-test design and some random assignment, Saxe, Gearhart, and Nasir (2001) also found positive effects of professional development interventions for mathematics teaching. They compared three sets of classrooms studying a unit on fractions. Two sets used the same reform curriculum, but the teachers in one group were randomly assigned to participate in an integrated professional development program while the teachers in the other group had no organized professional development, although they met regularly to discuss implementation of the curriculum. The professional development included a five day summer institute and 13 additional meetings. A third set of classroom teachers, not randomly assigned, used a traditional

curriculum. The study analyzed changes in conceptual understanding and computation. They found no difference between groups on the computation scale but did find systematic variation on the conceptual scale, with the reform group receiving professional development scoring substantially higher, more than a standard deviation, than the other two groups.

D. Summary

The evidence shows that policies aimed at influencing who becomes a teacher and what teachers do once they enter the classroom can change the teacher workforce and student outcomes. Wages influence teachers decisions; recruitment influences the pool of interested candidates; professional development, in some instances, can change teachers' behaviors and student outcomes. This said, we know little about the optimal design of teacher policies.

Across the board wage increases are extremely expensive. Among 3.5 million teachers staff classrooms in the United States, even a small across-the-board increase in wages is a huge expense. Targeted wage changes are more promising but difficult to design, given the many factors that influence a student's learning in a given year, the multitude of dimensions of learning that we care about (only some of which we measures), and the difficulty of design a reward system that benefits students equitably. Recruitment programs have dramatically changed the teaching force, particularly in large urban districts. Such approaches are likely to be a part of any effective comprehensive plan to improve teaching but they only affect the pool of new teachers (not the substantial number of individuals already teaching), and the evidence on how to select the best teachers from this growing pool of candidates is sparse. Finally, it is evident that professional development can improve student performance but that this professional development must be both intensive and targeted on specific tasks. Designing professional development that works on a large scale is a daunting task.

5. School Choice and Accountability

Schools may not be providing the best possible education for their students, not only because they lack resources or good teachers, but because they do not have the proper incentives. They may not have incentives to use their money wisely and they may be focusing on student outcomes that parents and communities do not value. Two sets of policies aim to realign incentives in order to improve opportunities for students: test-based accountability programs and market-based accountability programs.

In test-based accountability schemes, governments measure schools' achievement, judge whether they are successful or not, and attach a variety of rewards or sanctions to these judgments (Figlio and Ladd, 2008).[14] The best known of these policies is the Federal No Child Left Behind (NCLB) law of 2001, which required schools to make "adequate yearly progress" towards 100 percent student proficiency. But even before NCLB, many states and large cities had implemented accountability policies, which coexist with NCLB in states like California. Studies using pre-NCLB, cross-state variation in the timing of these state laws suggest some positive effects on test scores (Carnoy and Loeb, 2002; Hanushek and Raymond, 2005). Research within states has generally been limited in its ability to identify convincing comparison groups against which to compare the outcomes of students subjected to accountability provisions (Figlio and Ladd, 2008).[15] The strongest study uses variation in accountability pressures across schools in Florida and shows that schools facing greater pressure were more likely to implement

[14] Conceptually, measuring "success" involves estimating the causal effect of thousands of individual schools on test scores. In practical terms, accountability systems measure either the level of student performance in a given year and compare it a specified goal (e.g. the Federal No Child Left Behind law) or measure changes in schools' or students' performance between years (e.g. California's state accountability scheme). The dilemma in either case is that schools might be held accountable for variance in test score measures that is due to factors beyond schools' control (e.g. family poverty or randomness in test score fluctuations from year to year).

[15] One exception is a range of studies that examine effects of accountability pressures on schools judged to be failing in Florida. These studies, which use variants of discontinuity design, based on the formula for calculating "failure," suggest that test scores improved in these schools. See Rouse and Barrow (2008) and the citations therein.

a range of new instructional practices such as lengthening instructional time, focus more on low-performing students, and improving low-performing teachers. Moreover, improvements in student achievement in the schools are likely the result of these policy changes (Rouse, Hannaway, Goldhaber, and Figlio, 2007).

Market-based policies constitute a second approach to holding schools accountable. Broadly speaking, these policies enhance the ability of parents to choose a preferred public or private school. In so doing, they create incentives for school authorities to cater to parental preferences for certain features of schools and their students. There is already much choice through families' choice of residence and its neighborhood public school, which already creates competition (Hoxby, 2000; Rouse and Barrow, 2008). But, moving costs are high and not all parents have the resources and information needed to move to the neighborhood of their preferred school. Variants of other choice policies, such as private school vouchers and charter schools, are grafted onto this system of residential choice. The next two sections consider recent evidence on the effects of each policy on student outcomes.

A. Private School Vouchers

Private school vouchers are tuition coupons that students can redeem at a participating private school. In the few existing U.S. programs, voucher eligibility is typically restricted to small numbers of low-income students, and the participating schools are mostly Catholic (except on the occasions, such as the early phases of the Milwaukee program, when sectarian participation was restricted). The accompanying research has thus attempted to identify test score effects on low-income students who are offered or actually use a voucher to attend such

private schools. A separate literature, not considered here, considers how to estimate the general equilibrium effects of large school voucher plans.[16]

In the 1980s, when voucher plans were mostly hypothetical, authors used non-experimental methods and data to estimate the effect of Catholic school attendance on test scores. This literature, reviewed by McEwan (2000) and Neal (2002), showed no or very small effects on test scores, but more substantial effects on eventual high school attainment. Like the parallel debate on school resources, its results were inconclusive because of concerns that omitted variables like student motivation or ability were biasing estimates of private school effects.

As publicly- and privately-funded voucher programs were implemented in several U.S. cities, the evidence base improved. In 1990, Milwaukee's Parental Choice Program began offering vouchers of $2,446 (later increased) to low-income students for attendance at non-sectarian schools (Witte, 2002). Subsequent versions of the program included more students and private schools, but the best research was conducted in the program's early phase. Rouse (1998) compared achievement gains of students offered vouchers to gains of two comparison groups: a random sample of low-income students in Milwaukee Public Schools and, more compellingly, a group of unsuccessful applicants who were randomly denied admission to private schools. The results consistently suggested no statistically significant effects on reading scores, and small annual effects on math scores of no more than 0.11 standard deviations (Rouse and Barrow, 2008).

[16] The most compelling evidence on large-scale voucher plans is only available from countries like Chile that have actually implemented such plans (McEwan, 2001; Hsieh and Urquiola, 2006). For reviews of the wider literature on vouchers, see McEwan (2000), Zimmer and Bettinger (2008), and Rouse and Barrow (2008).

Privately-funded voucher programs have been implemented and evaluated with randomized experiments in several U.S. cities (Howell and Peterson, 2002; Rouse and Barrow, 2008). Most prominently, a New York City program offered $1,400 to poor children for private school attendance (if necessary, families were expected to contribute further towards private school tuition). Beginning in Fall 1997, a random subset of eligible applicants was offered vouchers and followed for three years by researchers. Two independent analyses found no effects of voucher offers on test scores after three years in the full sample of students (Mayer et al., 2002; Krueger and Zhu, 2004). The first study did find effects among the subsample of African-American students. Krueger and Zhu found that this result disappeared when using the full sample of data, and alternative methods of defining student race in the sample.

The best recent evidence of voucher's effects is from the randomized evaluation of a federally-funded voucher program in Washington, DC (the Opportunity Scholarship Program). The scholarships are worth up to $7,500 and can be used to cover tuition, fees, and transportation to any participating private school. As in New York City, the vouchers were restricted to poor students, and were awarded by lottery. After two years, the effect of the voucher offer on math scores is close to zero, and the reading estimates are 0.05-0.08 standard deviations, but none of these are statistically different from zero at the five percent level (Rouse and Barrow, 2008; Wolf et al., 2008).

B. Charter Schools

Charter schools are publicly-funded schools of choice that enjoy some degree of autonomy from local school authorities. They receive state or local funding based on the number of students that they attract. If they receive more applications than spaces, then students are usually admitted by lottery. Charter schools are not a homogeneous "treatment." In the 2007-08

school year, 40 states and the District of Columbia had enacted charter school laws with wide variation in charter authorization, finance, regulation, and accountability (Bifulco and Bulkley, 2008). Currently, about 4,100 charter schools enroll 1.2 million children (two percent of the total), although they are concentrated in a small number of states.[17]

The best research to date has focused on particular states or cities, and has followed one of two evaluation approaches, each with drawbacks. The first set of studies takes advantage of large samples of administrative data from states that track all students' test scores over time. The authors of these studies identify the subset of students that switch between public and charter schools, and compare their test scores, before and after, to the "non-switching" comparison group.[18] They are generally consistent in their findings, despite being conducted in Texas, North Carolina, Florida, and two large California cities (Hanushek et al., 2007; Bilfulco and Ladd, 2006; Sass, 2006; Zimmer and Buddin, 2006). Switching to charter schools often has negative effects, usually small, on student test scores (see Table 3). They tend to be largest when the charter school is relatively new, and closer to zero otherwise. The generalizability of these effects is uncertain, since they refer only to students that switch between grades, and not students who both start and complete their schooling in charter schools.

A second set of studies relies on the fact that charter schools are usually required to admit students by lottery when faced by excess demand. Hoxby and Rockoff (2004) compare the test score outcomes of students who won or lost in admissions lotteries at three Chicago charter schools. Overall, there were no statistically significant differences in reading or math scores between winners or losers, although this could mask some positive effects in earlier grades.

[17] National charter school data are regularly compiled by an advocacy group, the Center for Education Reform (http://www.edreform.com).
[18] Authors apply variants of student fixed-effects specifications. The exact specifications adopted by the authors differ, but the broad results are not sensitive to these decisions.

There is little national or state data collected on how many charter schools are over-subscribed, though even generous estimates conclude it is only a portion (McEwan and Olsen, 2007). By revealed preference of families, over-subscribed schools are perhaps the most effective of a city's charter schools. Thus, the Chicago results are surprising, but still broadly consistent with a more ambitious study that analyzed 194 admissions lotteries at 19 Chicago high schools (Cullen, Jacob, and Levitt, 2006). Though not charter schools, the high schools allow open enrollments in the same local schooling market. Despite evidence that participating families appear to choose better schools along a range of measures like test scores, the authors do not any evidence that lottery winners experience benefits on a wide range of achievement measures.

C. Summary

A premise of choice and accountability is that public schools use resources inefficiently. In the logic of test-based accountability systems, this inefficiency may arise from poor management or from schools aiming to produce outcomes other than test scores for students. In choice systems, the inefficiency could similarly be due to poor management or to schools aiming to produce outcomes that parents do not care as much about. In either case, there is under-production of student outcomes, which are presumably valued by parents and society. The two kinds of policies attempt to hold schools accountable and encourage better use of school resources.

Substantial recent research has asked whether these accountability programs, aimed to refocus the incentives facing school personnel, have improved student outcomes. The evidence on test-based accountability programs is mixed. However, it is clear that some systems can change school practices and, in turn, affect student learning (Rouse, Hannaway, Goldhaber, and Figlio, 2007). The evidence on the average effects of private school vouchers and charter

schools is even weaker. Few studies have shown positive effects that are statistically and socially meaningful. However, these programs, when implemented on a large scale, may provide incentives for innovation which, in the long run, can benefit schools and students.

6. Conclusions

We have known for some time that additional years of schooling are a good investment, but we know less about how to design education systems to use resources to maximize student outcomes. Fortunately, the volume and quality of research has accelerated in the past decade. This chapter's review focused on high-quality evidence on the impact and costs of interventions in three areas: direct resource investments, investments in the teacher workforce, and school choice and accountability.

Among direct investments, there is no consistent evidence that simply increasing expenditures will increase test scores, although such investments can increase achievement if used well. The research on class size reduction and intensive reading programs like Success for All provide evidence of the potential benefits of increased investments. In general, computer-assisted instruction is no panacea, though a recent study found it can be effective if coherently integrated with instructional goals and intensively applied. Among teacher policies, there is some evidence that across-the-board teacher wage increases can improve student outcomes, although this approach is quite costly. Evidence on targeted wage increase policies (like performance pay) is still sparse in the U.S. The mounting evidence is more consistent in suggesting that popular alternative routes for teacher recruitment, such as Teach for America, can raise test scores, at least in math, if they replace teachers with few formal qualifications. The vast literature on teacher professional development only suggests effects when the programs are intensive and targeted at improving specific student outcomes. Finally, a growing number of

randomized and natural experiments suggest zero or very small effects of receiving a private school voucher or gaining admission to a public school of choice.

This summary masks potentially large variation in the cost-effectiveness of the subset of "effective" programs and policies. Section 3 suggested that schools might have to invest upwards of $10,000 on class size reduction to obtain increases in test scores of at least 20 percent of a standard deviation in test scores. In other cases, such as Success for All or Teach for America, the same test score increases might be obtained for one-quarter the cost or less. Indeed, prior work has found, among a subset of effective interventions, that class size reduction is less cost-effective than others in raising test scores. These include computer-assisted instruction (Levin, Glass, and Meister, 1987) and investments in teacher resources (Grissmer et al., 2000).

These results might appear to suggest that class size reduction in not a worthwhile investment. However, this can only be judged by converting test score gains into a reasonable estimate of monetary benefits that can be weighed against costs. For example, Schanzenbach (2007) assumes that class size reduction raises test scores by 0.15 standard deviations and that a one standard deviation increase in test scores causes annual earnings to increase by 20 percent.[19] Under these assumptions, class size reduction shifts discounted annual earnings upward by three percent, using an age-earnings profile from the Current Population Survey. Weighed against the substantial costs of the Tennessee intervention, the intervention yields an internal rate of return of 4.8 percent, assuming no real wage growth. Krueger (2003) makes slightly different assumptions and finds an internal rate of return of 5.2 percent. Harris (2007) applies further

[19] The estimate is taken from Neal and Johnson (1996), who relate AFQT scores to subsequent earnings.

sensitivity analysis and finds that the internal rate of return does not fall below three percent, equal to a commonly applied discount rate.

The final chapter of this volume conducts a more careful cost-benefit comparison of class size reduction and other interventions. For the moment, however, the results illustrate that class size reduction—one of the *least* cost-effective education interventions—can at least pass a basic cost-benefit test (which only includes only a single category of benefits, private earnings). This implies substantial scope for identifying other economically reasonable investments in the quality of education. However, as the chapter's review suggested, the research literature still has far to go in separating the effective investments from the ineffective, and in thinking carefully about how to scale up pilot interventions.

References

Angrist, Joshua D. 2004. "American Education Research Changes Tack." *Oxford Review of Economic Policy* 20(2): 198-212.

Angist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In Orley Ashenfelter and David Card (Ed.), *Handbook of Labor Economics* (vol. 3A). Amsterdam: Elsevier.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2): 533-75.
Bacolod, Marigee. 2007. "Do Alternative Opportunities Matter? The Role of Female Labor Markets in the Decline of Teacher Quality." *Review of Economics and Statistics* 89(4): 737-751.

Bacolod, Marigee, John Dinardo, and Mireille Jacobson. 2008. "Beyond Incentives: Do Schools Use Accountability Rewards Productively?" Unpublished manuscript, University of California, Irvine, and University of Michigan.

Barnett, W. Steven. 1996. "Economics of School Reform: Three Promising Models." In Helen F. Ladd (Ed.), *Holding Schools Accountable: Performance-Based Reform in Education.* Washington, DC: Brookings Institution.

Barrow, Lisa, and Cecilia E. Rouse. 2005. "Causality, Causality, Causality: The View of Education Inputs and Outputs from Economics." Working Paper 2005-15. Chicago: Federal Reserve Bank of Chicago.

Barrow, Lisa, Lisa Markman, and Cecilia E. Rouse. 2008. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." Working Paper 14240. Cambridge, MA: National Bureau of Economic Research.

Belfield, Clive, and Henry M. Levin. 2007. *The Price We Pay: Economic and Social Consequences of Inadequate Education.* Washington, DC: Brookings Institution Press.

Bifulco, Robert, and Katrina Bulkley. 2008. "Charter Schools." In Helen F. Ladd and Edward B. Fiske (Ed.), *Handbook of Research in Education Finance and Policy*. New York: Routledge.

Bifulco, Robert, William Duncombe, and John Yinger. 2004. "Does Whole-School Reform Boost Student Performance? The Case of New York City." *Journal of Policy Analysis and Management* 24(1): 47-72.

Bifulco, Robert, and Helen F. Ladd. 2006. "The Impacts of Charter Schools on Student Achievement: Evidence from North Carolina." *Education Finance and Policy* 1(1): 50-90.

Bloom, Howard S. 2003. "Using 'Short' Interrupted Time-Series Analysis to Measure the Impacts of Whole-School Reforms: With Applications to a Study of Accelerated Schools." *Evaluation Review* 27(1): 3-49.

Borman, Geoffrey D., and J. Benson. Forthcoming. "A Randomized Field Trial of the Fast ForWord Language Computer-Based Training Program." *Educational Evaluation and Policy Analysis.*

Borman, Geoffrey D., and Gina M. Hewes. 2002. "The Long-Term Effects and Cost-Effectiveness of Success for All." *Educational Evaluation and Policy Analysis* 24(4): 243-66.

Borman, Geoffrey D., Robert E. Slavin, Alan C. K. Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers. 2007. "Final Reading Outcomes of the National Randomized Field Trial of Success for All." *American Educational Research Journal* 44(3): 701-31.

Boyd, Don, Grossman, Pamela, Lankford, Hamilton, Loeb, Susanna, and Wyckoff, James. 2006. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." *Education Finance and Policy, 1*(2).

Boyd, Don, Lankford, Hamilton, Loeb, Susanna, Jonah Rockoff, and Wyckoff, James. Forthcoming. "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools." *Journal of Policy Analysis and Management*.

Card, David. 1999. "The Causal Effect of Education on Earnings." In Orley Ashenfelter and David Card (Ed.), *Handbook of Labor Economics* (vol. 3A). Amsterdam: Elsevier.

Card, David, and A. Abigail Payne. 2002. "School Finance Reform, the Distribution of School Spending, and the Distribution of Student Test Scores." *Journal of Public Economics* 83: 49-82.

Carnoy, Martin, and Susanna Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educational Evaluation and Policy Analysis* 24(2): 305-331.

Carpenter, T. P., Fennema, E., Peterson, P.L., Chiang, C. P., & Loef, M. 1989. "Using knowledge of children's mathematics thinking in classroom teaching: An experimental study." *American Educational Research Journal* 26(4): 499–531.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools." *American Economic Review* 95(4): 1237-58.

Cook, T. D., Habib, F. N., Phillips, M., Settersten, R. A., Shagle, S. C., and 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): 543-97.

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): 535-97.

Corcoran, S, R. Schwab and W. Evans. 2004. Women, the Labor Market and the Declining Relative Quality of Teachers. *Journal of Policy Analysis and Management*. Vol 23. No. 3.

Corcoran, Sean P., and William N. Evans. 2008. "Equity, Adequacy and the Evolving State Role in Education Finance." In Helen F. Ladd and Edward B. Fiske (Ed.), *Handbook of Research in Education Finance and Policy*. New York: Routledge.

Cullen, Julie Berry, Brian A. Jacob, and Steven Levitt. 2006. "The Effect of School Choice on Participants: Evidence from Randomized Lotteries." *Econometrica* 74(5): 1191-230.

Cullen, Julie Berry, and Susanna Loeb. 2004. "School Finance Reform in Michigan: Evaluating Proposal A." In John Yinger (Ed.), *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity.* Cambridge, MA: MIT Press.

Darling-Hammond, Linda, Holtzman, Deborah J., Gatlin, Su Jin, and Heilig, Julian Vasquez (2005). "Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America and Teacher Effectiveness," *Education Policy Analysis Archives, 13*, No. 42, available online at http://epaa.asu.edu/epaa/v13n42/.

Decker, Paul T., Daniel P. Mayer, and Steven Glazerman. 2004. "The Effects of Teach For America on Students: Findings from a National Evaluation." MPR No. 8792-750. Mathematica Policy Research.

Downes, Thomas A., and David N. Figlio. 1998. "School Finance Reforms, Tax Limits, and Student Performance: Do Reforms Level-Up or Dumb Down?" Unpublished manuscript, Tufts University.

Downes, Thomas A., and David N. Figlio. 2008. "Tax and Expenditure Limits, School Finance and School Quality." In Helen F. Ladd and Edward B. Fiske (Ed.), *Handbook of Research in Education Finance and Policy*. New York: Routledge.

Dynarski, Mark, Roberto Agodini, Sheila Heaviside, Timothy Novak, Nancy Carey, Larissa Campuzano, Barbara Means, Robert Murphy, William Penuel, Hal Javitz, Deborah Emery, and Willow Sussex. 2007. "Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort." NCEE 2007-4005. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences.

Figlio, David F., and Helen F. Ladd. 2008. "School Accountability and Student Achievement." In Helen F. Ladd and Edward B. Fiske (Ed.), *Handbook of Research in Education Finance and Policy*. New York: Routledge.

Flanagan, Ann E., and Sheila E. Murray. 2004. "A Decade of Reform: The Impact of School Reform in Kentucky." In John Yinger (Ed.), *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity.* Cambridge, MA: MIT Press.

Glazerman, Steven, Daniel Mayer, and Paul Decker. 2006. "Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes." *Journal of Policy Analysis and Management* 25(1): 75-96.

Glewwe, Paul, Ilias, Nauman and Kremer, Michael (2003). Teacher Incentives. NBER Working Paper 9671.

Gordon, Nora. 2004. "Do Federal Grants Boost School Spending? Evidence from Title I." *Journal of Public Economics* 88(9-10): 1771-92.

Greenwald, Rob, Larry V. Hedges, and Richard D. Laine. 1996. "The Effect of School Resources on Student Achievement." *Review of Educational Research* 66(3): 361-96.

Grissmer, David, Ann Flanagan, Jennifer Kawata, and Stephanie Williamson. 2000. *Improving Student Achievement: What State NAEP Test Scores Tell Us.* Santa Monica, CA: RAND.

Guryan, Jonathan. 2003. "Does Money Matter? Estimates from Education Finance Reform in Massachusetts." Unpublished manuscript, University of Chicago.

Hansushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24(3): 1141-77.

Hansushek, Eric A. 2006. "School Resources." In Eric A. Hanushek and Finis Welch (Ed.), *Handbook of the Economics of Education* (vol. 2). Amsterdam: Elsevier.

Hanushek, Eric A., John F. Kain, Steven G. Rivkin, and Gregory F. Branch. 2007. "Charter school quality and parental decision making with school choice." *Journal of Public Economics* 91: 823-48.

Harris, Douglas N. 2007. "Class Size and School Size: Taking the Trade-Offs Seriously." In Tom Loveless and Frederick M. Hess, *Brookings Papers on Education Policy,* 137-61. Washington, DC: Brookings Institution Press.

Hill, Heather. 2007. "Learning in the Teaching Workforce." *The Future of Children* 17: 111-28

Hoxby, Caroline M. 2000a. "The Effects of Class Size on Student Achievement: New Evidence from Popular Variation." *Quarterly Journal of Economics* 115(4): 1239-85.

Hoxby, Caroline M. 2000b. "Does Competition Among Public Schools Benefits Students and Taxpayers?" *American Economic Review* 90(5): 1209-38.

Hoxby, Caroline M., and Jonah E. Rockoff. 2004. "The Impact of Charter Schools on Student Achievement." Unpublished manuscript, Harvard University and Columbia University.

Howell, William G., and Paul E. Peterson. 2002. *The Education Gap: Vouchers and Urban Schools.* Washington, DC: Brookings Institution Press.

Hsieh, Chang-Tai, and Miguel Urquiola. 2006. "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program." *Journal of Public Economics* 90: 1477-1503.

Jacob, Brian A., and Lars Lefgren. 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, Christopher, and Steven Rivkin. 2002. "What is the Tradeoff Between Smaller Classes and Teacher Quality?" Working Paper No. 9205. Cambridge, MA: National Bureau of Economic Research.

Kane, Thomas J., Rockoff, Jonah E. and Staiger, Douglas O. (2007). "Photo Finish: Certification Doesn't Guarantee a Winner," *Education Next, 7*, No. 1: 61–67

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

Krueger, Alan B. 2003. "Economic Considerations and Class Size." *Economic Journal* 113: F34-F63.

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

Krueger, Alan B., and Pei Zhu. 2004. "Another Look at the New York City School Voucher Experiment." *American Behavioral Scientist* 47(5): 658-98.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142(2): 675-97.

Levin, Henry M. 2002. "Issues in Designing Cost-Effectiveness Comparisons of Whole-School Reforms." In Henry M. Levin and Patrick J. McEwan (Ed.), *Cost-Effectiveness and Educational Policy* (Yearbook of the American Education Finance Association). Larchmont, NY: Eye on Education.

Levin, Henry M., Gene V. Glass, and Gail R. Meister. 1987. "Cost-Effectiveness of Computer-Assisted Instruction." *Evaluation Review* 11(1): 50-72.

Levin, Henry M., and Patrick J. McEwan. 2001. *Cost-Effectiveness Analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Leuven, Edwin, Mikael Lindahl, Hessel Oosterbeek, and Dinand Webbink. 2007. "The Effect of Extra Funding for Disadvantaged Pupils on Achievement." *Review of Economics and Statistics* 89(4): 721-36.

Loeb, Susanna and Marianne E. Page. 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, Susanna, and Katharine Strunk. 2003. "The Contribution of Administrative and Experimental Data to Education Policy Research." *National Tax Journal* 56(2): 415-38.

Mayer, Daniel P., Paul E. Peterson, David E. Myers, Christina Clark Tuttle, and William G. Howell. 2002. "School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program." MPR Reference No. 8404-045. Mathematica Policy Research.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2): 698-714.

McEwan, Patrick J. 2000. "The Potential Impact of Large-Scale Voucher Programs." *Review of Educational Research* 70(2): 103-49.

McEwan, Patrick J. 2001. "The Effectiveness of Public, Catholic, and Non-Religious Private Schools in Chile's Voucher System." *Education Economics* 9(2): 103-28.

McEwan, Patrick J., and Robert Olsen. 2007. "Admissions Lotteries in Charter Schools." Unpublished manuscript, Wellesley College and Urban Institute.

Muralidharan, Karthik and Sundararaman, Venkatesh (2006). Teacher Incentives in Developing Countries: Experimental Evidence from India, Harvard University Working Paper.

National Center for Education Statistics (NCES). 2008. *Digest of Education Statistics 2007*. Washington, DC: U.S. Department of Education.

Neal, Derek. 2002. "How Vouchers Could Change the Market for Education." *Journal of Economic Perspectives* 16(4): 25-44.

Papke, Leslie E. 2005. "The Effects of Spending on Test Pass Rates: Evidence from Michigan." *Journal of Public Economics* 89: 821-39.

Raymond, Margaret, Fletcher, Stephen H., and Luque, Javier. (2001). *Teach for America: An Evaluation of Teacher Differences and Student Outcomes in Houston, Texas* (Stanford, CA: The

Hoover Institute, Center for Research on Education Outcomes [CREDO], 2001), available online at http://credo.stanford. edu/downloads/tfa.pdf.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417-458.

Rouse, Cecilia Elena. 1998. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics* 113(2): 553-602.

Rouse, Cecilia Elena, and Alan B. Krueger. 2004. "Putting Computerized Instruction to the Test: A Randomized Evaluation of a 'Scientifically Based' Reading Program." *Economics of Education Review* 23: 323-38.

Rouse, Cecilia Elena, and Lisa Barrow. 2008. "School Vouchers and Student Achievement: Recent Evidence, Remaining Questions." Occasional Paper No. 163. New York: NCSPE, Teachers College.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio (2007). "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure," NBER Working Paper 13681.

Sass, Tim R. 2006. "Charter Schools and Student Achievement in Florida." *Education Finance and Policy* 1(1): 91-122.

Saxe, G. B., Gearhart, M., and Nasir, N. S. 2001. "Enhancing Students' Understanding of Mathematics: A Study of Three Contrasting Approaches to Professional Support." *Journal of Mathematics Teacher Education* 4: 55–79.

Schanzenbach, Diane Whitmore. 2007. "What Have Researchers Learned from Project STAR?" In Tom Loveless and Frederick M. Hess, *Brookings Papers on Education Policy*, 205-28. Washington, DC: Brookings Institution Press.

Schwartz, Amy Ellen, Leanna Stiefel, and Dae Yeop Kim. 2004. "The Impact of School Reform on Student Performance: Evidence from the New York Network for School Renewal Project." *Journal of Human Resources* 39(2): 500-22.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Teach for America, "In Strong Job Market, Record Number of Graduating Seniors Apply to Teach for America" (June 1, 2006), available online at http://www.teachforamerica.org/newsroom/documents/TeachForAmerica_News_20060601.html .

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113(485): F3-F33.

Urquiola, Miguel. 2006. "Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia." *Review of Economics and Statistics* 88(1): 171-7.

Urquiola, Miguel, and Eric Verhoogen. Forthcoming. "Class Size Caps, Sorting, and the Regression-Discontinuity Design." *American Economic Review.*

van der Klaauw, Wilbert. 2008. "Breaking the Link Between Poverty and Low Student Achievement: An Evaluation of Title I." *Journal of Econometrics* 142: 731-56.

What Works Clearinghouse. 2007. "Technical Details of WWC-Conducted Computations." Washington, DC: Institute of Education Sciences, What Works Clearinghouse.

Witte, John F. 2000. *The Market Approach to Education: An Analysis of America's First Voucher Program.* Princeton, NJ: Princeton University Press.

Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, Nada Eissa, and Marsha Silverberg. 2008. "Evaluation of the DC Opportunity Scholarship Program: Impacts After Two Years." NCEE 2008-4023. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences.

Xu, Zeyu, Jane Hannaway, and Colin Taylor. 2008. "Making a Difference: The Effects of TFA in High Schools." CALDER Working Paper.

Yoon, Kwang Suk, Duncan, Teresa, Lee, Silvia Wen-Yu, Scarloss, Beth, and Shapley, Kathy L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement.* National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education. Report REL 2007–No. 033

Zimmer, Ron, and Eric P. Bettinger. 2008. "Beyond the Rhetoric: Surveying the Evidence on Vouchers and Tax Credits." In Helen F. Ladd and Edward B. Fiske (Ed.), *Handbook of Research in Education Finance and Policy*. New York: Routledge.

Zimmer, Ron, and Richard Buddin. 2006. "Charter School Performance in Two Large Urban Districts." *Journal of Urban Economics* 60: 307-326.

Table 1: The effects of direct resource investments

Study	Intervention	Grades(s) (length) of intervention	Research Design	Sample; Year(s) of Intervention	Outcomes (posttest grade)	Effects
van der Klaauw (2008)	Title I funding allocations (~5% of expenditures)	K-12	Discontinuity assignment based on poverty	New York City public schools; 1993, 1997, 2001	School pass rates in reading and math	n.s. or negative effects, but offsetting effects on school expenditures
Bacolod, DiNardo, and Jacobson (2008)	Unrestricted school awards (~5% of expenditures)	K-12	Discontinuity assignment based on composite test scores	California public schools; 2000-01	School-mean composite test score	n.s.
Guryan (2003)	Added state funding	K-12	Discontinuity assignment based on funding formula variables	Massachusetts public school districts; 1994-96	District-mean math (4^{th}) District-mean reading (4^{th}) District-mean math (8^{th}) District-mean reading (8^{th})	.15σ per $500 (1996) .06σ per $500 (1996) Positive but non-robust effects in 8^{th}
Schanzenbach (2007)	"Small" classes (13-17) vs. "regular" classes (22-25)	K (4 yrs.)	Randomization of students/teachers within schools	79 Tennessee schools; students; 1985-89	Composite test (3^{rd}) Composite test (8^{th}) Took college entrance exam (change in probability)	.15σ (full sample) .24σ (black students) .12σ (white students) n.s .02 (full sample) .05 (black students) n.s. (white students)
Borman et al. (2007)	Success for All reading program	K (3 yrs.)	Randomization of schools	41 schools in 11 states (2001-2004)	Multiple reading tests (2^{nd})	.21-.36σ
Dynarski et al. (2007)	16 technology products for reading/math instruction	1, 4, 6 (1 yr.)	Randomization of products/training to teachers within schools	132 schools, 439 teachers (2004-2005)	Multiple reading and math tests	n.s.
Barrow, Markman, and Rouse (2008)	Computer-assisted math instructional package	8-10 (1 yr.)	Randomization to class periods within schools	17 schools, 61 teachers (2004-2005)	Algebra test	.17σ

Note: n.s. indicates not statistically significant at 5%. Reported estimates from Schanzenbach (2007), Borman et al. (2007), and Barrow et al. (2008) are intent-to-treat effects.

Table 2: The effects of investments in teachers

Study	Intervention	Grades (length) of intervention	Research Design	Sample; Year(s) of Intervention	Outcomes (posttest grade)	Effects
Loeb and Page (2000)	Across-the-board wage increase	All teachers	State-level difference-in-difference analysis with IV	1970-1990	High school graduation, college enrollment	3-4% drop in dropouts and 1-2 percent college enrollment increase for 10% wage increase
Glazerman et.al. (2006)	Teachers selected and trained by TFA		Randomized assignment of students to TFA or non-TFA teachers within grades	6 cities, 17 elementary schools, 100 classrooms	Reading Math	n.s .15σ
Boyd et al. (2005)	Teachers selected and trained by TFA		School fixed effects	NYC student-level data, grades 4-8, 1998-04	Reading Math	-.03σ n.s
Xu et al. (2008)	Teachers selected and trained by TFA teachers		Student fixed effects	North Carolina student-level data, high school, 2000-06	Math and Science	.07σ
Jacob and Lefgren (2004)	Externally-provided training to schools (17 firms)		Discontinuity assignment of training subsidies to low-performing schools	Chicago Public Schools, grades 3-6, 1996-99	Reading Math	n.s. n.s.
Carpenter et. al (1989)	Professional development workshop		Experiment	First grade teachers	Math	.5 standard deviations in complex addition and subtraction (positive but not significant overall)
Saxe, Gearhart, and Nasir (2001)	Professional development		Experiment	Upper elementary math teachers	Math	1.5 standard deviations on the conceptual scale, n.s. in computation

Note: n.s. indicates not statistically significant at 5%.

Table 3: The effects of vouchers and charter schools

Study	Intervention	Grades (length) of intervention	Research Design	Sample; Year(s) of Intervention	Outcomes	Effects
Rouse (1998); Rouse and Barrow (2008)	Offer of private school vouchers (up to $2,985 in 1993)	K-8 (annual gains)	Comparison groups of unsuccessful applicants and random sample of Milwaukee public students	Milwaukee; 3163-8751 students (depending on comparison group (1990-94)	Reading Math	n.s. (annually) n.s. to .11σ (annually)
Krueger and Zhu (2004)	Offer of private school vouchers (up to $1400)	K-4 (3 yrs.)	Random assignment of vouchers to eligible (poor) applicants	NYC; 2080 students (1997-98)	Reading Math	n.s. n.s. (Small, non-robust effects for black subsample)
Wolf et al. (2008)	Offer of private school vouchers (up to $7500)	K-12 (2 yrs.)	Random assignment of vouchers to eligible (poor) applicants	Washington, DC; 2308 students; 2004-06	Reading Math	n.s. n.s.
Hanushek et al. (2007)	Student switching between public and charter school	4-8	Student fixed effects	TX administrative data (1996-2002)	Composite reading and math	-.32σ to n.s.
Bifulco and Ladd (2006)	Student switching between public and charter school	3-8	Student fixed effects	NC administrative data; 1996-2002	Reading Math	-.18σ to -.06σ -.31σ to -.08σ
Sass (2006)	Student switching between public and charter school	3-10	Student fixed effects	FL administrative data; 1999-2003	Reading Math	-04.σ to -.01σ -.08σ to -.02σ
Zimmer and Buddin (2006)	Student switching between public and charter school	Elementary and secondary	Student fixed effects	Los Angeles and San Diego administrative data; 1997-2002	Reading Math	-2.1 to n.s. (elementary) -1.2 to 1.5 (secondary) -5.0 to n.s. (elementary) -1.7 to 1.3 (secondary)
Hoxby and Rockoff (2004)	Offer of place in 1 of 3 Chicago International Charter Schools	1-8 (1 yr.)	Lottery admissions in school-by-grade blocks	Chicago, 2668 students	Reading Math	n.s. n.s. (some positive and significant effects for younger students)

Note: n.s. indicates not statistically significant at 5%. Reported estimates from Rouse (1998), Krueger and Zhu (2004), Wolf et al. (2008), and Hoxby and Rockoff (2004) are intent-to-treat effects. Effects in Zimmer and Buddin (2006) are reported in test score percentiles.