

This paper was published as

Betts, Julian R. and Paul Hill, (2010), “Through a Glass Darkly: An Introduction to Issues in Measuring the Quality of Charter Schools”, Chapter 1 in Paul T. Hill and Julian R. Betts (Eds.), **Taking Measure of Charter Schools: Better Assessments, Better Policymaking, Better Schools**, Lanham, MD: Rowman & Littlefield Publishers, Inc.

Through a Glass Darkly: An Introduction to Issues in Measuring the Quality of Charter Schools

by Julian R. Betts and Paul Hill

Charter schools are semi-autonomous public schools that receive renewable charters to operate, typically from a host school district or university. Charter schools typically do not hew to the local district's collective bargaining agreement, nor do they strictly follow the district's curriculum and pedagogical approach. In return for this semi-independence, charter schools are accountable to the host district for academic results. The host district has the option of closing down a charter or deciding not to renew its charter agreement. Even casual observers of education policy know that charter schools are controversial. Since 2004 there has been a widely covered debate between researchers disagreeing on the meaning of data on charter school performance.¹ There have been major legislative fights over whether to permit additional charter schools in California and New York. Some school district leaders have denounced charter schools for depleting their budgets.

None of these events has stopped the growth in numbers of parents applying to charter schools. Nor has it stopped local public education leaders in New York, Chicago, DC, and elsewhere from embracing chartering as a way to provide new options for children. State governments forced to take over collapsing school districts in Pennsylvania, California, and Louisiana have also turned to chartering.

Despite the controversy, nobody seriously expects charter schools to go away. Opponents can denounce and oppose charter schools, but they have no success in reducing the numbers of charter schools or killing the special state laws on which they are based. Even teacher unions, which in Ohio and elsewhere have filed lawsuits hoping to block the growth of charter, have also decided to start charter schools of their own.

Charter schools, or something very much like them, are here to stay. This is so, not because charter schools have always been proven superior to other forms of public school or because proponents have always won the fights described above. They haven't. However, charter schools offer something that public school systems, parents, and teachers need: a way to experiment with alternative ways of teaching, motivating

¹ See, for example, Carnoy et al (2005) and Henig (2008).

students, organizing schools, using technology, and employing teachers. Even in localities like Chicago and Philadelphia, where charter schools are plentiful but state law limits their further growth, district leaders are creating contract and partnership arrangements that look a lot like charters.

By exempting schools from many regulations and collective bargaining, chartering opens up possibilities for new uses of public money, teaching talent, student work, time, and technology. As we have seen, chartering also creates possibilities for failed experiments and big mistakes. On balance, however, Americans are willing to give charter schools a chance. Indeed a 2007 nationwide poll found that 60% of respondents stated that they favored charter schools, compared to 35% who opposed and 5% who were undecided. Support was even higher, at 63%, among public school parents. (Rose and Gallup, 2007)

That does not mean that citizens in general, or even charter supporters, are satisfied that the value of charter schools has been proven.

Like many other public policies, charter school laws were enacted without a great deal of thought about how their effects would be measured and judged. Proponents assumed that charter schools would perform so well that their superiority could be seen with the naked eye. Children would benefit so dramatically that parents would demand more and more charter schools and elected officials would become strong supporters.

Opponents, also fearing that charters would be visibly effective, prepared objections of the “yes...but” variety. Opponents expected to attack charter schools based on discrimination in admissions and other abuses, not performance. Everyone was surprised by how difficult it was to assess charter school performance. Somehow, it had been assumed, children would be tested and scores could be compared. But often children weren’t tested, or comparison was difficult because the students at a given school changed over time, making comparisons of overall trends in average achievement of dubious value. Other problems have been that charter schools did not use the same tests as district schools, and school records didn’t tell much about students’ characteristics and prior educational experience.

Opponents were surprised that charter schools did not discriminate against poor and minority students; to the contrary they served such students in disproportionate

numbers. Everyone was surprised to find that many students entered charter schools with serious educational deficits. It was therefore difficult to set definite expectations for those students' performance, and therefore not obvious from the simple snapshots of average school performance how much students gained from attending charter schools.

Thus, supporters and opponents – not to mention neutral public officials and citizens – faced unexpected challenges in judging charter school performance.

Researchers sought to assess charter school performance by comparing the test scores of students attending charter schools with students in regular public schools. But study results – whether they showed positive or negative effects for charter schools – were subject to withering criticism. It proved extremely difficult to find a credible comparison group against which to compare charter students' scores. Few non-charter schools served exactly the same mix of students as did charter schools, and researchers could never be sure that a given group of students attending regular public schools was a perfect match for students attending charter schools.

The 2004 dustup over charter school research illustrates how far we are from having the unambiguous evidence on charter school performance. Nobody can “win” the debate about whether students attending charter schools benefit, because the data for good analysis just isn't available on a nationwide basis, although it is indeed available in certain cities and a handful of states.

The Charter School Consensus Panel

The National Charter School Research Project (NCSRP) was founded in 2004 to take a hard look at charter schools and become a trustworthy source, both of evidence about charter school performance and of ideas about how that performance can be improved. This book is a result of one of NCSRP's first initiatives, which was to assemble a consensus panel of top scholars to review charter school research and suggest ways parents, educators, and policymakers could get valid evidence about charter school performance.²

² Consensus Panel members include Julian Betts, University of California San Diego, Dominic Brewer, University of Southern California, Anthony Bryk, Stanford University, Dan Goldhaber, University of Washington, Laura Hamilton, RAND Corporation, Jeffrey Henig, Columbia University, Paul Hill, University of Washington, Susanna Loeb, Stanford University, and Patrick McEwan. Wellesley College.

The Consensus Panel's first product was a White Paper, *Key Issues in Studying Charter School Achievement: A Review and Suggestions for National Guidelines*, published in May 2006. It considered the strengths and weaknesses of different methods for estimating how much students learn as a consequence of attending charter schools. As the White paper showed, there are three basic approaches to estimating a charter school's benefits to students:

- Comparing the scores of students attending charter schools with those of students who applied to the same schools but did not get in because all the seats were taken.
- Comparing individual students' test scores before and after entering charter schools, in order to judge whether students' learning rates were higher or lower in charter than in non-charter schools.³
- Comparing scores for students in charter versus non-charter schools, matched on the basis of students' income, race, and other educationally relevant factors (e.g. home language, immigrant status, handicapping conditions).

As the White Paper explained, in theory the first method, comparing scores of charter school students with others who applied to the same schools but lost in a lottery, is best because it compares students who are on average identical in all ways (including their desire to enroll in a charter school) and are distinguished only by the luck of the draw.

The second method is also very good because it uses individual students as their own controls; scores are compared before and after a student transfers between a public school and a charter school.

By contrast, the third assessment method is tricky because it involves comparing different students. It can produce valid or invalid results--depending on how well researchers match up students in charter and regular public schools. Comparisons of groups with big differences in income, race, parents' education, and ESL status can be highly misleading. Valid comparisons can be difficult even if the researcher controls all relevant student characteristics. For example, if the students in a charter school have unusually committed parents or unusually high prior achievement levels, demographic matching will ignore key factors and almost certainly make the charter school look good for reasons other than the effectiveness of its program. The same point can be made in the

³ The full text of the White Paper explains the different ways data collected for a study using this method can be analyzed.

opposite direction. A charter school may have a disproportionate number of children who left regular public schools because they were doing much worse than others of their same economic or racial group. Students remaining in regular public schools were not motivated in the same way and are therefore different.

A particularly weak, yet common, version of the third method is simply to compare average test scores in a given year across schools with rudimentary or even no controls for student characteristics across schools. Such snapshots tell us nothing about growth in achievement.

Whether one method or another can be used to assess a particular charter school or group of schools depends on local conditions and the availability of data. The first method can only be used in a locality where charter schools have lotteries with waiting lists. The second method can only be used in localities where annual test scores are kept for all students, including those who transfer between charter and district-run public schools.

The White Paper noted that most of the charter school research done to date is limited by the quality of data available. Researchers are often stuck with databases that make valid comparisons difficult. The Consensus Panel suggested that readers of a study consider the quality of data on which it is based, asking questions like:

--Does it include test scores for multiple years or just one year? A one-year snapshot can give a misleading result if, for example, students in one kind of school (charter or regular public) had higher average scores before the year in which the snapshot was taken. Though more studies use one-year snapshots than any other method, they cannot lead to results as definitive, no matter how large a database they draw from or how sophisticated the analysis. In short, a study that does not control for the academic history of the student in some way is likely to go awry.

--Does the study include detailed information about the students in charter schools? Weak data on student attributes—which can make dissimilar students look alike and similar students look different—can wreck efforts to compare performance of students from different schools. 4

⁴ However, two methods reduce the need for detailed student characteristics. Lotteries, by definition, ensure that on average lottery losers and winners will have about the same

--Does the analysis include good information about factors correlated with school effectiveness? How long, for example, has the school been operating (new charter schools struggle much more than older ones), is the school financially stable, and what is the turnover rate among teachers and school leaders?

--Have students in charter schools--and students to whom they are compared--been tested in the same way? When charter school students take one test and the district-run school students to whom they are compared take another, gaps in outcomes can be due to differences in the tests rather than to school quality.

No single research method is perfect, and it is seldom possible to get ideal test scores or complete information about schools and students. Any rigorous study, for example, would try to control for the proportions of low-income of students in charter vs. regular public schools, but many charter schools do not participate in the free or reduced price lunch program, a common proxy for low-income status. As a result, counts of students in the lunch program may provide rough estimates of student poverty in regular public schools but seriously underestimate the number of low-income families in charter schools. Some researchers have no alternative but to use free and reduced price lunch counts as their measure of low-income status--but the results must then be interpreted very carefully.

Every study, in short, includes some compromises. And researchers and readers must be clear about how those compromises limit the applicability of findings in charter schools.

Beyond the White Paper

Though the White Paper was well received for clarifying the debate about charter school performance, members of the Consensus Panel thought their work had just begun. It is one thing to say what the best research methods are, and quite another to show that they make a difference in study outcomes. Similarly, it is one thing to urge researchers to use

characteristics, because they have been assigned to the two groups by a flip of the coin. The second method we mentioned – using student “fixed effects” to compare individual students’ performance gains when in charter vs. regular schools. In this method, we do not have to compare one student to another.

the results of admissions lotteries to find control groups, but another to say how to identify a lottery that would produce a truly randomly selected control group.

The White Paper left many issues unresolved, including how researchers might use outcome measures other than test scores (e.g. students persistence in school and ability to succeed at the next higher level of education), and how studies might factor in richer information about school programs and teacher qualities. The White Paper also urged researchers to consider how elected officials, funders, and others use information about charter school performance, without explaining how those parties actually used data.

Thus, immediately after publishing the White Paper the Consensus Panel committed to looking much more deeply into a number of these issues. Individual Consensus panel members took responsibility for many of the needed analyses, and other scholars with special knowledge were also invited to contribute.

This result, this book, is in two parts: The first part focuses on how to improve estimates of charter schools' performance, especially their benefits to students who attend them; the second part suggests how policymakers can learn more about charter schools and make better use of evidence.⁵

Part One: Improving Research On Charter Schools

Julian Betts and colleagues lead off in Chapter 2 with a new analysis of student achievement results from San Diego charter schools. Local data allow them to analyze student achievement data using alternative methods, some relatively crude (e.g. comparing averages scores of students in different schools), and some highly sophisticated (comparing test score trajectories of students before and after enrolling in charter schools). They show not only that more sophisticated methods lead to richer results, but also that better methods can produce a totally different message about charter performance. In their San Diego data, naïve analyses whose methods and results resemble some of the cruder studies done on national databases, prove negatively biased against finding positive charter school outcomes. It is possible that in other local contexts the bias would run in the other direction. Indeed, in their re-analysis of experimental results

⁵ The Consensus Panel has also published a media guide to help reporters understand the limitations of charter school studies. See Appendix 1.

at one San Diego school, naïve models overstate rather than understate the true causal effect of attending the given charter. They find that the more sophisticated models can produce the same sign and similar coefficients on the estimated effect of attending on student achievement. But the simpler models that do not incorporate value-added are seriously biased.

Laura Hamilton and Brian Stecher discuss non-test outcomes in Chapter 3. They remind readers that test scores provide incomplete information about school effectiveness and fail to capture all of the outcomes that parents and educators truly care about – students’ completion of courses, graduation, ability to gain admission and succeed at higher levels of education, find productive work, and act as effective citizens. They suggest other milestone indicators that might be used to supplement test scores, (e.g. attendance, teacher quality, stability of enrollment) and also identify more authentic long-term measures that can more fully represent the consequences of charter schooling. Examples of the latter include ability to take college courses without remediation, postsecondary degree attainment, employment, earnings, and civic values. Hamilton and Stecher conclude with recommendations on how non-test data might be collected and how researchers and local officials can set priorities for data gathering and dissemination in order to obtain and report on the most important of these hard-to-measure outcomes.

In Chapter 4, Julian Betts reviews the existing charter school achievement studies, using the criteria for sound methods first developed by the Consensus Panel. The chapter also summarizes and analyzes results of the first studies to assess charter schools’ long term effects on student attendance, persistence in school, graduation, and college attendance. Betts concludes that the preponderance of evidence on achievement suggests that charter schools are outperforming traditional public schools, but that there are important variations and some locations and subject areas/grades (e.g. math in high school) in which charter schools appear to be underperforming. There is now a small literature that models outcomes apart from test scores. These studies are small in number but provide hints that charter schools may increase the probability of graduating from high school and the probability of attending college, while reducing the number of student disciplinary actions. Much more needs to be done on all research fronts.

In Chapter 5 Julian Betts addresses a problem that vexes all efforts to measure the effectiveness of charter schools – whether the children whose parents choose charter schools are so different from the children in regular public schools that straightforward comparisons are impossible. Betts shows that in theoretical models low-income families choose charter schools because of their perceived greater quality (compared to the public school options available to them) but that many factors can reduce disadvantaged children’s attendance charter schools, even when school operators intend to serve such children. Low-income families, he notes, are especially sensitive to transportation issues, so school location is extremely important. Some families are discouraged from applying by requirements for parental participation in school activities. Though he concludes that many charter schools intend to serve disadvantaged minority (especially African-American) populations, their actual locations and recruitment practices can favor slightly more educated and economically secure black families. Betts concludes with suggestions about how researchers can test and control for schools’ selectivity bias.

Patrick McEwan and Robert Olsen take a careful look at charter school lotteries in Chapter 6, and explore implications for policy and research. The authors describe why some charter school conduct lotteries and some do not, how they conduct them, and why all lotteries are not created equal. In addition, they explore the effects that lotteries may have on equal access and stratification in public schools. Finally, they discuss lottery-based studies of charter school effectiveness, which compare student outcomes between lottery winners, who are admitted to charter schools by random chance, to lottery losers. McEwan and Olsen conclude that lottery-based studies have enormous potential when the lottery details are well understood, but that they have important limitations as well. The authors conclude by noting that requirements to conduct lotteries in public may increase the transparency of lotteries, and that requirements to report lottery results might help provide opportunities for both monitoring—to ensure the lotteries are truly random—and research to exploit the natural experiments that charter school lotteries provide.

In Chapter 7, Paul Hill and Lydia Rainey suggest that charter school maturation should be a factor in studies of school performance. As they show, many charter schools are relatively new; moreover, charter schools consistently have lower scores in the first years after they open than at later times. Hill and Rainey apply literature from business

and broader education research to show that new schools are likely to have growing pains that will depress performance early. They therefore suggest that studies of charter performance should distinguish schools in their first three years from older charters. They also consider the possibility of using school maturation measures as control factors in research, or as leading indicators to alert families and authorizing agencies to possible trouble. They conclude, however, that schools mature at such different rates and in such different sequences that it is impossible to create a model of normal maturation.

In Chapter 8, Dominic Brewer and June Ahn review what is known about charter school teachers and consider how data on teachers might be used to explain differences in school performance. They show that charter school teachers are younger and lower-paid than teachers in regular public schools serving similar students. Though many charter teachers have educational backgrounds generally similar to those of regular public school teachers, teacher qualifications are highly variable in charters. Similarly, teacher turnover is high and variable; moreover the significance of turnover depends on whether charter schools have definite strategies for managing it (or simply scramble constantly to fill classrooms). Charter school teachers generally work longer hours than their public school counterparts, but schools differ between those that routinely cause teacher burnout and those that help teachers sustain needed levels of effort over a long time. To date, links between teacher factors and charter school performance have not been shown. But as Brewer and Ahn conclude, teacher factors might prove to be important in explaining variations in charter school outcomes. To the degree possible, studies of charter school performance should account for differences in teacher qualifications, satisfaction, and stability of employment.

Part 2: How Policymakers Can Make Better Use of Evidence

In Chapter 9, Robin Lake and Larry Angel review studies done on charter schools within individual states. They show that state studies are highly variable in quality, but that they include many of the best studies based on following students from before to after they enter charter schools. States that keep records on individual students and have score data for every year a student is tested are in position to sponsor very sound studies. State legislators and their staff members can be excessively demanding and reasonable in

turn. State legislation mandating charter studies can pose questions that cannot be answered given the low quality of data kept by their states; yet, officials generally know that definitive evidence about charter schools is not forthcoming soon. Angel and Lake suggest ways legislatures can both discipline their requests for results and upgrade the quality of data so that the answers they need will ultimately become available.

Jeffrey Henig examines controversies over the political uses of charter school research in Chapter 10 and draws lessons about how policy-makers might use better research as it becomes available. He provides both bad news and good news about the use of research results in policymaking. The bad news is that policymakers are not well equipped to tell the difference between strong and weak studies, especially of a relatively new phenomenon like charter schools, and that institutions that could help them interpret the existing research are not serving that function as well as they should.. Under such circumstances, elected officials are more likely to use studies to buttress pre-established positions than to weigh options. The good news is that the weight of research can accumulate over time, especially when the quality of studies increases and their results converge on key points. Henig concludes with recommendations to researchers who want their work used and trusted, emphasizing the need to frame the consequences of research modestly and to avoid oversimplification of findings in search of headlines.

In the final chapter 11, Julian Betts and Paul Hill distill lessons from the all the Consensus Panel's work. They suggest ways states and localities can improve the quality of data on which charter school studies are based and trace some of the ways charter school research influences policy. Consistent with Jeffrey Henig's conclusions, they find no hard link between research and public policy, though elected officials care enough about studies to fund them and to cite results in support of their positions. However, research results have powerful effects inside the charter school movement. Funders, school operators, and government authorizing agencies are using research aggressively, responding even to mixed study results with new quality standards for charter schools to meet. Finally, Betts and Hill note that charter school research is starting to influence the data kept on regular public schools and the ways individual schools are assessed. No Child Left Behind, for example, will require districts to measure and judge the performance of all their schools.

The ultimate result, they argue, will be that all public schools will be assessed on the same student performance standards that now apply almost exclusively to charters. However, charter schools still face much more severe consequences for perceived failures than do traditional public schools. It is true that No Child Left Behind, with its stipulation that traditional public schools that repeatedly fail to make adequate yearly progress must be reconstituted, re-organized, or converted into charter schools, brings stronger accountability to non-charter schools. But even these interventions still pale compared to the power that a chartering authority, usually a school district, has to shut down a charter school completely simply by refusing to renew its charter.

Student achievement tests have been used for decades, and administrators and researchers have long claimed that they were evaluating school performance. However, until recently test scores had few real consequences. Public schools were not closed if their scores were low, or expanded or rewarded in any particular way if their scores were high. In the 1990s, standards-based reforms enacted by all but a few states adopted the rhetoric of performance-based accountability, but no state followed through completely. Now test scores can have real consequences, at least for charter schools. schools that admit students by choice.

Now that the stakes have been raised, data and methods that once looked acceptable prove inadequate. As the Consensus Panel has shown, the wrong data, or the right data used wrongly, can lead to unwarranted conclusions school success and failure. There are now real incentives for hard thinking and careful use of data. Charter school research has improved slowly but steadily over just the last five years, and it will continue improving.

References

Carnoy, Martin, Rebecca Jacobson, Lawrence Mishel, and Richard Rothstein, 2008. , *The Charter School Dustup, Examining the Evidence on Enrollment and Achievement*, New York, Teachers College Press 2005.

Henig, Jeffrey, *Spin Cycle: How Research is Used in Policy Debates: The Case of Charter Schools*, New York: Russell Sage Foundation/Century Foundation, 2008.

Rose, Lowell C. and Alec M. Gallup. 2007. The 39th Annual Phi Delta Kappa/ Gallup Poll Of the Public's Attitudes Toward The Public Schools, Phi Delta Kappan, September, 89:1. Also downloadable from http://www.pdkintl.org/kappan/k_v89/k0709pol.htm.