

The Constructive Use of Existing Data and Research for Evaluating Charter Schools

Gary Miron
The Evaluation Center
Western Michigan University

Paper prepared for the Symposium on the Use of School-Level Data for Evaluating Federal Education Programs, December 8-9, 2005, Washington, D.C.

Introduction

In academic circles, considerable attention is given to sophisticated approaches and methods for working with less than desirable data like those contained in the School-Level State Assessment Database (SSASD). In practice, however, many of the sophisticated approaches are not applied due to cost and time restrictions and sometimes because methods elaborated on paper are undermined by unforeseen lapses in data quality that do not become apparent until data are sorted and prepared for analysis. Some of these unforeseen data limitations include (i) missing or incomplete background data for experimental and/or control groups, (ii) large proportions of students in either the experimental and/or control groups that do not take the test, switch schools, or are retained and retake the same test in consecutive years, and (iii) changes in state assessment tests or norm groups that can occur far too frequently over time.

In our work at The Evaluation Center we have been contracted to evaluate a wide range of school programs and education reforms. Interestingly, every time we work with student achievement data, the design and analytical strategy changes depending on the quality and consistency of achievement data and overall project time span and funding. In this paper, I have drawn upon the evaluations of charter school reforms we have conducted. These will illustrate the wide range of design alternatives that can be used. Particular attention will be given to the various strategies that can be used to capture gain or change scores using school- or group-level data because this is the what is available for use in the SSASD.

The examples covered in this paper are grouped by design type. We generally group the student achievement studies in charter schools in the following design categories: cross-sectional, successive cohorts, same cohorts, similar groups of students, matched students, and random assignment. There are – of course – many variations of these designs since they differ in the nature and number of controls used, and they vary in terms of the overall scope of the studies (i.e., number of years, numbers of grades covered, numbers of subjects included, etc.).

Given that the findings regarding student achievement in charter schools are hotly contested and garner substantial attention, and given the overall importance of student achievement for the

Copyright © 2005 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.

Opinions and statements included in the paper are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Planning Committee for the Symposium on the Use of School-Level Data in Evaluating Federal Education Programs or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

survival and growth of these schools, it seems that a focus on charter schools will provide a suitable context for discussing study designs.

Charter schools have been around for over a decade now and the body of research on charter schools has clearly improved over time. Initially, most of the writing on charter schools was rather rhetorical and not empirical. The published research on these schools in the mid-1990s was largely theoretical in nature or else focused on start-up issues and the degree to which these schools were innovative. By the end of the 1990s, more of the research and evaluations of charter schools addressed student achievement. Improvements have occurred in the research of charter schools over time for the following reasons:

- ❑ Charter schools have operated for more years so there are more schools to include in the analysis with multiple years of data to track.
- ❑ As charter schools grow in size, more schools pass the threshold at which their data can be made public (e.g., some states like Ohio wait until a school's third year before it will publicly report data. Most states also do not post data for subgroups when there are too few testtakers).
- ❑ State assessment systems are expanding and improving over time. All states are adding more grades to be tested, and several states are making greater efforts to equate results from year to year. Also a number of states are developing and allowing access to individual student data with unique identifiers that allow researchers to link students from year to year.
- ❑ Finally, the overall sophistication of the charter school research has improved over time as researchers adapt and create more rigorous methods of measuring growth in charter schools with less than desirable data.

Over the past decade, a number of state agencies and foundations have contracted with The Evaluation Center to evaluate the implementation and impact of charter school reforms. The state level evaluations have all relied on available state assessment systems to provide data on outcomes. In other words, the general source of data for these studies is the same data that feeds into the SSASD. In total, we have conducted 9 large scale evaluations of charter school reforms in 6 different states as well as a multi-state study of schools operated by Edison Schools Inc. (many of these Edison schools were charter schools). Our own work has not included studies in the weakest category of designs (i.e., cross sectional), nor has our work included any randomly controlled experimental designs.¹ While a few of our studies have found charter schools to be performing poorly in Michigan and Ohio, we have also conducted evaluations that found charter schools in Connecticut and Delaware to be outperforming (i.e., outgaining) students in traditional public schools.

In the following sections, the examples that are discussed are grouped based on the overall design category, starting with cross-sectional studies and then progressing to increasingly more

¹ In our statewide evaluation of charter schools in Illinois, we did attempt to replicate random assignment by comparing students who entered the charter schools with students who applied but did not gain access through the lottery process. In other words, the control group would be drawn from the students who did not gain entry to over-subscribed schools. In the end, we had to scrap this approach because the waiting lists were too small in numbers or were not sufficiently audited and maintained over time.

rigorous designs. A few analytical approaches are discussed in greater detail because they offer promising means of evaluating student achievement data in the absence of student- or individual level data. In the closing section of the paper, means of synthesizing the findings from diverse studies of student achievement by weighing them based on the quality of their overall design are discussed. So too are recommendations for strengthening and increasing the utility of the SSASD.

Cross-Sectional Studies of Student Achievement

Interestingly, some of the most widely discussed and debated studies of student achievement in charter schools have been those that were the weakest in terms of design. In the autumn of 2004, the AFT (2004) released a report that included a cross-sectional analysis comparing charter schools to traditional public schools using the NAEP data. This report received front page coverage in the New York Times and was billed as having sweeping implications. A researcher from Harvard University (Hoxby, 2004) responded by releasing her own cross-sectional study a few days later that claimed to be based upon a comparison of 99 percent of all charter schools in the nation with their neighboring noncharter public schools. The findings from the two studies had opposing conclusions and stirred considerable debate. The National Center for Educational Statistics (NCES, 2005) released a more complete analysis of the NAEP data in the spring of 2005 which had similar conclusions to the AFT report (see NCES, 2005). Hoxby responded to the NCES study by releasing a slightly modified version of her earlier cross-sectional analysis which again concluded that charter schools were making substantial gains compared to traditional public schools.

While these cross-sectional studies provided a snapshot picture of how charter schools were performing at a single point in time, both studies attempted to provide some controls for differences in populations. The studies based on the NAEP data blocked and compared subgroups of students by ethnicity, family income (i.e., free or reduced lunch status) and special education status. The Hoxby study made an assumption that neighboring schools would be demographically similar.

Aside from the multi-state studies noted above, cross-sectional designs have often been used by the media and advocacy groups to assemble quick and easy comparisons of student achievement in charter schools relative to traditional public schools. A few states have also relied on cross-sectional designs to evaluate their charter schools. One noteworthy example of this is the Colorado Department of Education which has been preparing cross-sectional comparisons for several years without calculating change or gain scores for successive or same groups over time. Given the expansion and improvement of the Colorado state assessment system, one expects more sophisticated designs will be used in the future, possibly even a design that will be based on student-level results instead of school-level results.

Successive Cohorts

Most of the available studies of student achievement in charter schools fall within the category of successive cohorts design. Basically, the design allows for comparison of successive groups of students over time and for comparisons to be drawn between charter and noncharter public schools. This is a common approach as only a few states allow access to individual student data researchers

can only compare groups over time. Also, until recent years, state assessment systems tested and reported data at only a few grade levels (e.g., grades 4, 8, and 10) which made following same groups over time more difficult. The analysis of successive cohorts means that different or successive groups of students are compared at the same grade level and in the same subject area. For example, Reading results for fourth graders in 2001 are compared to Reading results for fourth graders in 2002. The key assumption with this design is that students in successive years will have more exposure to the treatment (i.e., charter school). Another key assumption is that the schools attract and enroll students with similar background characteristics each year so that changes in student performance are not affected by shifting demographics over time. Unfortunately, these assumptions do not hold true since mobility rates in charter schools can be very high, and when schools are fully implemented they may attract students with differing characteristics than they did during the start-up phase.

Studies in this category vary considerably depending on whether and how well they control for demographic differences between the charter and noncharter public schools that are being compared. Some studies use blocking of data and draw comparisons between subgroups of students based on ethnic background, FRL status, Title 1 funding status, special education status, etc. Other studies have controlled for differences using regression analyses and consider many or all of the same demographic characteristics.

Two rather unique approaches that analyze successive cohorts over time include residual gains analysis and odds-ratio analysis. In the following two sections these approaches are briefly described.

*Residual Gains Approach*²

In our 3 year evaluation of Pennsylvania charter schools (see Miron, Nelson, & Risely, 2002) we analyzed changes in residuals over time. This was based on the state assessment data that captured mean scale scores that provided a more sensitive measure and allowed comparison from year to year. Only school- or group-level data was available and only students in Grades 5, 8, and 11 took the state assessment at the time of our evaluation.

The key finding of this evaluation was that Pennsylvania charter schools appeared to be having a modestly positive influence on student achievement. Yet, a simple examination of scores on the Pennsylvania System of School Assessment (PSSA) suggested that most charter schools scored well below the state average. In our full technical report we explained how both statements can be true. The answer lay in the distinction between score levels and score gain. In short, Pennsylvania charter schools appeared to be attracting students with lower-than-average achievement levels and producing small relative gains in their achievement levels.

Since achievement *gains* are much less correlated with student background factors than student achievement levels, they provide a good indicator of school effectiveness. At the time of our evaluation (2002), students were assessed at three grade levels only (5, 8, and 11). Thus, instead of observing a single group of students (e.g., fifth graders) as they progressed into the sixth grade and

² The example outlined in this section is based on the methodology section of the final report for the evaluation of Pennsylvania charter schools (Miron, Nelson, & Risely, 2002).

beyond, we were restricted to observing the performance of consecutive groups of students. By themselves, trends in PSSA performance do not allow us to distinguish score changes that are due to school effectiveness from those that are due to changes in student composition.

To estimate the charter school effect, we developed a set of statistical “filters” that subtract most of the changes in student composition over time in the charter schools. The remaining portion of the score changes provides a reasonable (though not foolproof) estimate of school effectiveness. While calculating the filtered scores requires statistical techniques (described in great detail in the technical report), the basic idea is relatively simple. The filters work by comparing each charter school with a set of demographically and geographically similar noncharter public schools. Instead of focusing on absolute levels of PSSA scores, the filtered scores focus on the differences between each charter school and a specially selected comparison group created using regression analyses. Variables used in the filters were obtained from state and federal databases and included income, race, special education status, urbanicity, PSSA participation rates, and school enrollment.

Inasmuch as the comparison schools were similar to charter schools in most relevant respects save for not being a charter school, the filtered (difference) scores provided the best approximation of the charter school effect given the available data.

Another advantage of the filtered scores is that they have a straightforward interpretation. Since the filtered scores represent the difference between a charter school and its specially selected comparison group, a score of zero indicates that the charter school is performing exactly on par with its comparison group. Unlike most test metrics, filtered or residual scores can take negative values, indicating that the charter school’s performance is below that of its comparison group. For example, a filtered score of -50 indicates that the average student in a charter school scored 50 points lower on the PSSA than predicted or relative to the average student in the school’s comparison group. By contrast, a filtered score of 78 indicates that the average student in a charter school scored 78 points higher than the average student in the school’s comparison group.

As discussed above, changes in scores over time provide a better estimate of value added than a snapshot from a single point in time. Charter school gains in residual scores over time indicate that the average student score is catching up with the average student in the school’s comparison group. Similarly, declines in a charter school’s residual scores suggests that the average student in the charter school is falling behind the average student in the school’s comparison group. Appendix A includes a snapshot of the findings from this evaluation.

The regression analysis allows for statistical controls that enable the analyst to compare each charter school’s score with demographically similar schools. Assume, as an illustration, that the analyst wants to create comparison groups based only on the concentration of low income students in a school. In this case, the analyst would simply regress the charter school test scores against income for all noncharter public schools. Explained in intuitive terms, the regression procedure simply finds the line (mathematical function) that best relates income to PSSA scores. Mathematically, this entails finding the line that minimizes the distance between each data point in two-dimensional space and the regression line.

Figure 1 provides a graphical example. The top line running from northwest to southeast is the regression line for all noncharter public schools. The regression line can be viewed as the set of predicted pass rates for each level of income. Alternatively, the regression line may be viewed as the set of mean PSSA scores for comparison schools at each level of income.

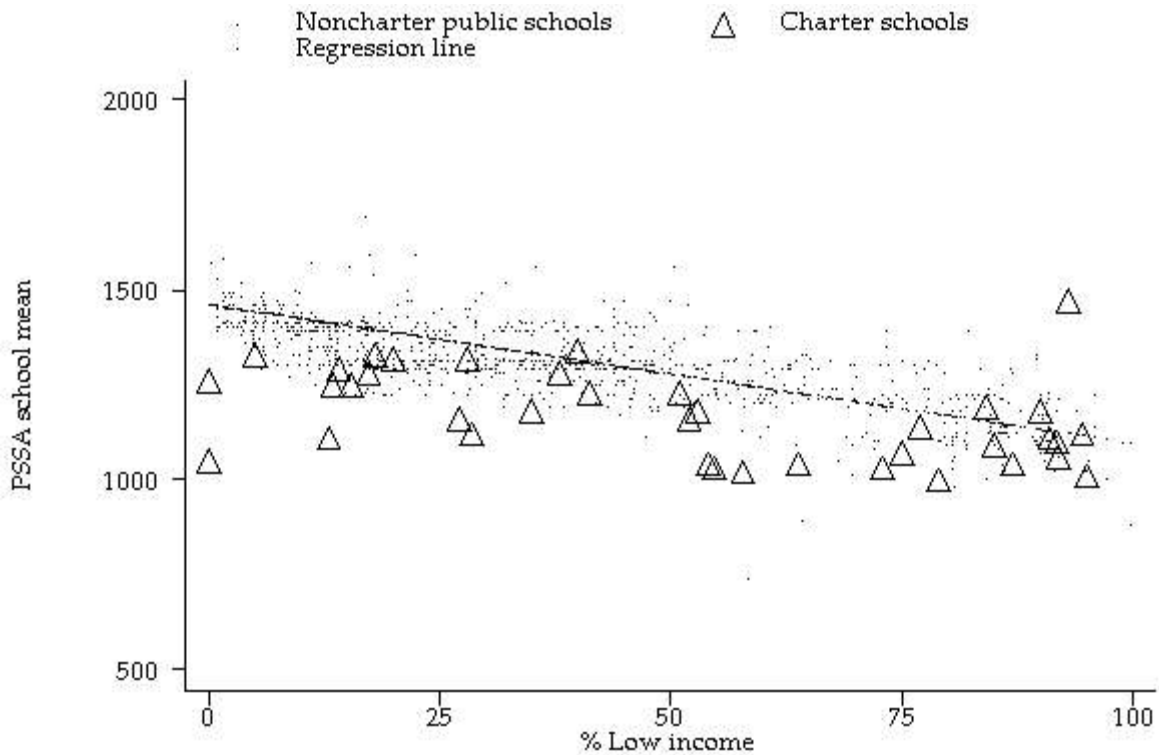


Figure 1. Illustration of Using Regression to Calculate Residual PSSA Scores

To get the residual score, we calculate the difference between the observed pass rate and the pass rate predicted by the model. Put another way, it is the difference between the charter school's pass rate and that of its demographically matched comparison group. The procedure for generating comparisons based on more than one demographic variable requires a related, though more complex, approach called multivariate regression. The basic idea, however, remains the same. Graphically, the multivariate regression model is extended into multidimensional space, with an additional dimension for each demographic variable added.

In any regression model, the accuracy of the predicted values (and thereby the residuals) depends upon the choice of independent variables. For this evaluation we relied on standard models of student achievement found in the production function literature and elsewhere. The final model employed in the regression analysis modeled PSSA scores as a function of family income, race, concentration of special education students, enrollment, PSSA participation rate, and urbanicity. In all instances, the regression models were estimated only for noncharter public schools in districts sponsoring charter schools. This allowed us to control for charter-noncharter differences that are not fully captured by the other variables. Appendix A includes some of the findings from this study which illustrates the usefulness and appropriateness of this approach to measuring student achievement in charter schools with less than desirable group level data.

*Odds-Ratio Analysis*³

In our evaluation of student achievement in schools operated by Edison Schools Inc., we employed several different statistical methods and analyses which, taken together, provide a composite picture of student performance on standardized tests at Edison schools. In this section, a description of how we used odds ratio analysis to make cross state comparisons using school level results on the differing state assessment tests is included. Odds ratio analysis is a strategy more commonly used in epidemiology. It was suitable for this particular study because it allowed us to make cross-state comparisons while relying on student learning outcomes from varying state assessment tests. Limitations in state assessment data meant that we could only compare successive groups or cohorts of students over time.

The odds ratio analysis we conducted examined student learning outcomes within a successive cohort study by analyzing the collapsed ordinal responses (pass/fail categories) on the state tests. A cohort study is when subjects are selected before they are exposed to possible determinants of interest (i.e., being in a charter school), and their exposure to possible determinants of interest (i.e., “the charter effect,” or “the Edison effect”) are then recorded along with the outcome (i.e., passing or failing the state test or in other words, meeting or exceeding state standards vs. not meeting state standards). The critical design factor in a cohort study is the comparability (similarity) of the two groups at the beginning of the time period under study. If the two groups are similar, then an observed association between being in an Edison school and passing (or failing) a component of the state test can be reasonably defended. However, if the two groups are not similar, then any observed association between being in an Edison school and passing (or failing) a component of the state test may or may not be truly a function of attending an Edison school.

Although there are several possible ways to define passing, we opted to define passing and failing as specified by each state assessment system. For example, if the criterion-referenced state assessment test is scored along a 4-point scale (Level 1 [lowest] to Level 4 [highest]) and the state criteria for passing is a score of 3 or 4 then we collapsed level 4 into level 3 and level 1 into level 2 to define passing and failing respectively. It should be noted that this reclassification could mask some important gains evidenced by the students in either the Edison charter school or the comparison group. We constructed the 2x2 tables for these analyses in such a way to represent the relative odds for a student to fail a component of the state test.

The odds ratio (OR) (McNeil, 1996) is defined as $OR = ad/bc$ and represents the proportion of students who fail the test in the treatment school relative to the proportion of students failing the test in the comparison school. An odds ratio can take values from zero to positive infinity. Interpretation of an OR is straightforward. An OR value of 1.00 represents equal odds for failing (or passing) relative to the comparison group. Values from 0.00 to 1.00 are representative of a “protective” effect; that is, the odds of failing are lower in the Edison school. Values greater than 1.00 would represent increasing odds for failing the test if enrolled in the Edison school. As with any point estimate, a $(1-\alpha)$ confidence interval (CI) needs to be constructed for accurate interpretation. Thus, if the CI around the OR includes 1.00, the conventional interpretation would

³ The example outlined in this section is based on the methodology section of the final report for the evaluation of student achievement in schools operated by Edison Schools Inc. (Miron & Applegate, 2000).

be that there is no statistically significant difference in the relative failing rate between the two schools (i.e., if the CI included 1.00, there is no statistically significant difference). However, if the CI does not include 1.00, the OR is generally interpreted as statistically significant, either representing a statistically significant protective effect or a statistically significant increase in the odds for failing the test. Due to the truncated nature of the sampling distribution of the OR, the standard error of the OR is calculated based on the natural logarithm of the OR, similar to converting a correlation to a Fisher's Z before constructing a $(1-\alpha)$ confidence interval around a correlation. The standard error of the natural log of the OR is

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{n_1 - a} + \frac{1}{b} + \frac{1}{n_2 - b}}$$

This approach allowed us to measure the nature and relative size of trends over time in the Edison schools relative to state and local district trends. Appendix B includes sample findings using the odds ratio analysis from one of the charter schools operated by Edison Schools Inc. that was included in this study.

Same Cohorts

Same cohort studies refer to those instances where we are able to follow same groups of students over time. For example, comparing grade 4 reading in 2003 to grade 5 Reading in 2004. As states roll out their state assessment systems to cover more grades, it is increasingly possible to follow same cohorts rather than successive cohorts. In the examples in the previous section, test data were not available for consecutive grades so it was difficult to track the same group or cohort of students to the next test event which could be between 2 and 4 years later. Calculating gain scores on same cohorts is preferable to consecutive cohorts although there is still the possibility that students leave or are added to the group over time which cannot be controlled for without individual student data. Gain scores for same cohorts can be used in both the residual gains or odds-ratio analyses described in the previous section. The charter school students that have followed same cohorts are few in number and include the work we did in our Connecticut evaluations (Miron & Horn, 2002; Miron 2005).

Matched Students ⁴

While the cross-sectional, successive cohort, and same cohort design categories are all based on group or school level data, there are an increasing number of charter school studies that are based on individual student data. When researchers or evaluators can get their hands on individual student data, it is almost always the case that they can link this to demographic data. Likewise, it is almost

⁴ The example outlined in this section is based on the methodology section of the final report for the evaluation of Delaware charter schools (Miron, 2004).

always the case that they can link student scores over time to measure gains or changes in performance. This allows analysts to match charter and noncharter students on a few or several demographic characteristics and then track their relative performance over time. The first study of student achievement in charter schools that used a matched student design was completed in 2001 for Arizona (see Solmon, Paark, & Garcia, 2001).

In this section, I will describe our longitudinal study from Delaware that allowed us to match every charter school student to a demographically similar noncharter school student. Gain scores were calculated over time after controlling for previous performance levels. First, the source and nature of the data are described. Then a description of how we designed and compiled charter school and comparison groups in separate panels is included. Finally, a description and justification for the analytical strategy as well as some of its limitations is included.

In addition to its extensive warehousing of school level data, the Delaware Department of Education has an advanced performance data system that yields and tracks data for all students in the state. A data set was provided to us by the Department of Education with test data in two subject areas from the past 7 years. This dataset included both students in charter schools and students in traditional public schools. Identifying information was removed and replaced with unique identifier codes that allowed us to link students from year to year. The scope and nature of these data allowed us to use a matched student design to examine the impact that charter schools were having on student learning. The matched student design is a quasi-experimental design in which students in the experimental group (i.e., charter schools) are matched according to all relevant background and demographic indicators with students in the control group (i.e., traditional public schools). Students are followed over time and we track and compare relative gains.

About the assessment instrument. Data for the analyses are from the Delaware Student Testing Program (DSTP), which is the statewide assessment program. The DSTP is used to measure how well students are prepared relative to the Delaware Content Standards in English language arts, mathematics, science, and social studies. Our analyses focused on reading and math because data were available for more years and the tests in these subject areas had both scale scores and normal curve equivalents as outcome measures.

Results from the test are reported at various levels, including the state, district, school, and individual student. Individual student data are carefully protected by the state, and obtaining access to these data involved an application and permission process. The data obtained for our analyses had unique identifiers which allowed us to track and link students from year to year. The results are reported by grade and subject area and the measures used include both scaled score results on the DSTP and the normal curve equivalent (NCE) scores on the SAT-9. A number of items from the SAT-9 are incorporated in the DSTP math and reading tests (not the writing component) so that equivalent scores can be calculated for the SAT-9.

Panel definition. The goal of our panel definition was to create a random sample of noncharter students who were demographically matched with charter school students that spanned the greatest number of DSTP assessments. Multiple panel designs were considered. Our aim was to use a panel design with three data points; however, this resulted in too few students with valid test scores at all three data points. This was due to student mobility and the fact that many charter

schools did not exist or had limited grade range in the early years of the reform. Development of the six panels (A - F) illustrated in the table below began with the most current DSTP assessment year (either 2003 or 2004) and looked back in time to the previous DSTP assessment. Thus, we were able to build three panel pairs that examined longitudinal growth from third to fifth grade, fifth to eighth grade, and eighth to tenth grade. As can be seen in the table below, the panel sample size in the more recent assessment years and at younger grade levels is greater than in the earlier and older assessments, reflecting an increasing enrollment trend for charter schools.

Description of the Panels

Panel	Total Number of Charter School Students in Analysis			Year of DSTP Data With Test Grades Highlighted in Bold				
	Math	Reading	Writing	2000	2001	2002	2003	2004
A	515	491	516			3rd	4th	5th
B	428	411	427		3th	4th	5th	6th
C	328	316	328	4th	5th	6th	7th	8th
D	295	293	284	5th	6th	7th	8th	9th
E	221	211	222	6th	7th	8th	9th	10th
F	180	179	181	7th	8th	9th	10th	

A panel was created by merging one DSTP subject area (reading or math) with the demographic data and selecting subjects who had valid test data in the two years selected for the panel and who were in the target grade in the last panel year, e.g., grade 5 in 2004 in Panel A. Once the appropriate population of students was selected, e.g., the above condition, the matching and random selection processes were undertaken.

Charter students were matched with noncharter students on four demographic characteristics: gender, ethnicity, Title I status, and FRL status. It is important to note that charter school status was defined by where a student was enrolled in the final DSPT assessment for that panel. According to the codebook supplied by DOE, there were five coding levels for ethnicity and two each for gender, Title I, and FRL. Thus, there were 40 different demographic strata for matching. We also considered matching on special education status (two levels) and limited English proficiency (two levels), but this resulted in 160 possible demographic combinations. There was almost no variability in these last two demographic variables, so they were not considered further. After the 40 demographic strata were defined, the total panel population was broken down among the 40 strata for charter schools and noncharter schools. After the panel population was stratified, demographically matched samples could be randomly drawn from each strata.

Analytical strategy. To address the central research question (i.e., is there a difference in achievement between students attending charter schools vs. students attending noncharter schools), an analysis of covariance (ANCOVA) was conducted on the last DSTP assessment with the previous DSTP assessment score as the covariate. Separate ANCOVA analyses were examined for DSTP scaled score and SAT-9 NCE for the reading and math assessments.

The use of the previous DSTP as the covariate served as a statistical matching procedure where the means on the last DSTP assessment for each group (charter and noncharter) are adjusted to what they would be if the two groups had scored equally on the previous DSTP assessment. Thus, using the previous DSTP assessment is a statistical control for previous achievement level; as such, the evaluative question directly addressed by the ANCOVA is “Is enrollment in a charter school associated with higher DSTP mean assessment scores in math and reading than enrollment in a noncharter school after adjustment for previous DSTP assessment performance?” A portion of the actual findings are included in Appendix C to help illustrate the utility of this approach.

Randomly Controlled Experiment

Conducting a randomly controlled experiment with charter schools involves randomly assigning students to either a charter or noncharter public school. This is a very complex and difficult study to conduct because it requires the recruitment of families to participate and abide by the decision to enroll in the school they are assigned (either charter or noncharter). In reality, families that are dissatisfied with their local public school and apply for a place in a charter school are likely to seek other options than their local public school if they do not get a place at the charter school. For example, they are likely to apply again for the charter school in the subsequent year or they may seek to move their child to a private school. This is only one example of why such studies are complicated and costly. Attempts have been made to use students who apply but do not gain access to an oversubscribed charter school as a control group. This represents random assignment but may mask other factors such as families who give up their place because of lack of transportation. One study in a small group of Chicago charter schools was conducted by Hoxby and Rockoff (2004) which found that charter schools outperformed students who did not gain a place through the lottery system and had to continue in district schools. While questions have arisen regarding the scope and incomplete nature of the technical report for this study, the U.S. Department of Education has funded a rather large and expensive randomly controlled study of student achievement in charter schools which is being conducted by Mathematica. Results from this longitudinal study will begin to be reported in 2006.

Synthesis of Findings Across Studies

Policymakers are looking for definitive answers regarding the success of education programs and new school reforms such as charter schools. Although charter schools have been operating for more than a decade and even though these schools were expected to be highly accountable schools, we still do not have a definitive answer for policymakers regarding whether or not these schools are

improving student achievement. Some argue that we will not have a definitive answer until we can conduct a randomized experiment. Others, including myself, have attempted to provide an answer to policymakers by summarizing or synthesizing the existing body of research on the topic. Ideally, these efforts to synthesize the research would use a metaanalytical approach (see Glass, 1976) or at least a best-evidence approach (see Slavin, 1986), however, these approaches require a more sophisticated set of studies that calculate and report effect sizes.

Given the limitations in data available for research on charter schools, it is not possible to calculate and synthesize effect sizes. Therefore, it has been difficult to provide systematic and rigorous summaries of the research. Together with colleagues at The Evaluation Center, we have made an effort to provide a synthesis of the growing body of research on charter schools to provide answers for policymakers (see Miron & Nelson, 2001; 2004).⁵ Our approach has been to construct a picture from the diverse studies available, even though many of these studies use relatively weak study designs. Each new study on student achievement in charter schools brings the picture into greater focus. Our approach to synthesizing the research involves the four steps listed below:

1. Setting inclusion and exclusion criteria to determine which studies will be considered.
2. Next, the selected studies were categorized by the nature of the impact on student achievement (i.e., Very negative, Negative, Mixed, Positive, and Strongly positive). Finally, the studies were weighted by the quality of the design. The weights given to studies also consider scope of the studies in terms of number of years, proportion of schools included in the analyses, and the number of grades and subjects included.
3. The final step involved weighing and synthesizing the results to calculate average state and national performance levels.

Our results have shown that when synthesizing the research across the nation the results are mixed or very slightly negative toward charter schools. Large differences exist by states however. While our initial synthesis in 2001 included 12 studies, an update of this report considering studies completed by the spring of 2003 (see Miron & Nelson, 2004) included 17 studies. A new update of our synthesis of the research on charter school performance is due out in the late spring of 2006. This synthesis will include consider and weigh the findings from more than 30 studies of student achievement in charter schools that now exist. Figure 2 below provides a tentative illustration of how the findings vary by state, by the nature of their impact (i.e., either positive or negative), and by the overall quality of the study.

⁵ There have been a few other attempts to synthesize the research on student achievement. Back in 2001, RAND compiled a summary of the research on student achievement in charter schools (see Gill et al., 2001). Because they set the bar so high for the studies to be included, only 3 studies were actually considered and summarized. More recent summaries have been prepared by Hassel (2005); Carnoy et.al. (2005), and Hill (2006). The later three examples have grouped studies depending on whether they were positive or negative and in some cases they are group by design type although no effort has been made to weigh and synthesize the results across studies.

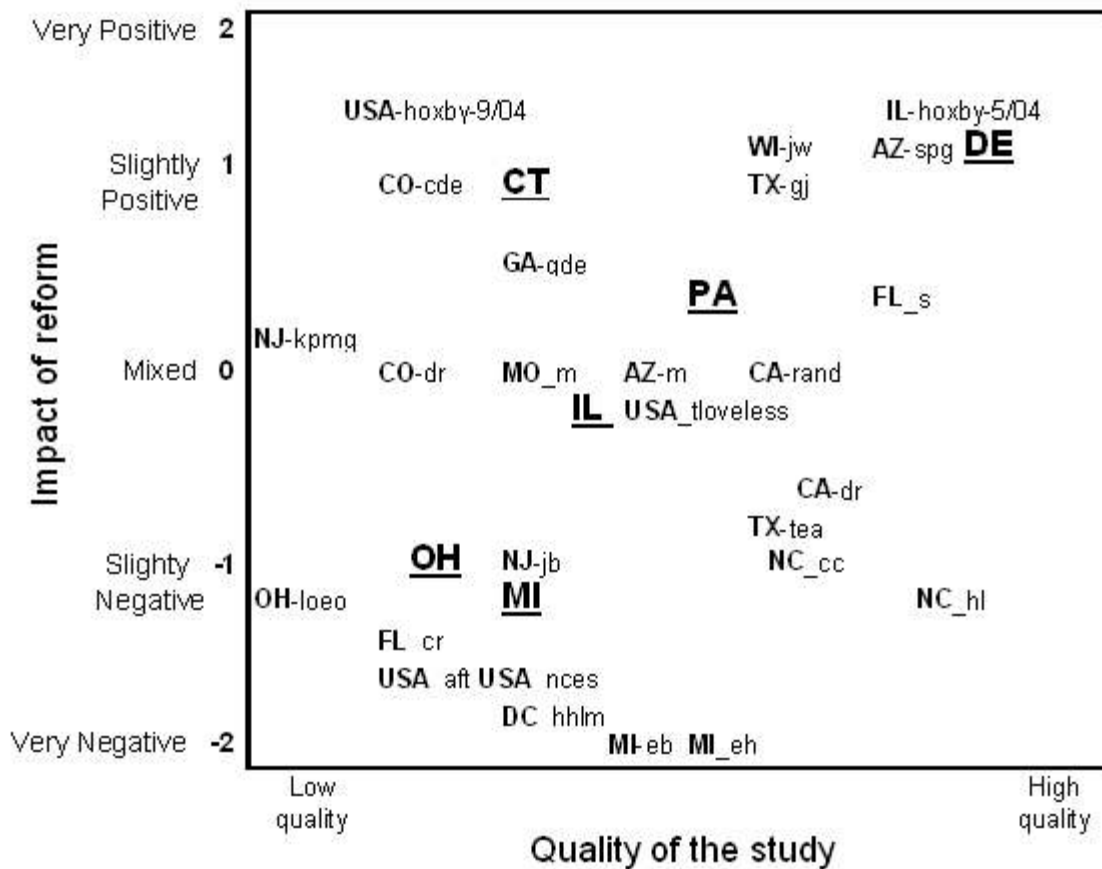


Figure 2. Quality and Impact Ratings for State and National Studies of Student Achievement in Charter Schools

Note: This figure provides an illustration of estimated impact and quality ratings for 32 studies completed during the past 5 years. Those studies completed by The Evaluation Center are highlighted in bold and underlined text.

Reviewing and Weighing Studies of Student Achievement

In this section, a number of key questions are raised that need to be considered when reviewing and weighing studies of student achievement.

Study Design

Key questions to address:

What is the design for the study?

Is the study longitudinal or cross-sectional?

Does the study use individual student data or group level data?

Below is a general list of possible study designs that are rank ordered from strongest (top) to weakest (bottom). Randomized experimental designs (#1) are widely considered the “gold standard.” These designs are often difficult to use in educational settings. If they are not implemented correctly or if they do not meet the necessary conditions for use, the strength of the design will be undermined. In other words, this is the strongest design when properly used.

Longitudinal designs	Student-level data	1. Randomized experiment
		2. Matched groups
		3. Similar groups with controls
		4. Similar groups with no controls
		5. Dissimilar groups
	School-level data	6. Cohorts with controls
		7. Cohorts with no controls
		8. Successive cohorts with controls
		9. Successive cohorts with no controls
		10. Cross-sectional with controls
		11. Cross-sectional with no controls

A quasi-experimental design often used in education is matched groups (#2). This refers to studies where students are matched with a comparison group that is similar in all identifiable respects.

Design #3 draws comparisons with students who are not similar on all key characteristics although statistical adjustments are made to account for differences that may exist. Design #4 also compares similar groups but without statistical controls. Design #5 compares dissimilar groups over time without statistical controls. Designs 3, 4, & 5 are not common in practice since when researchers can get their hands on individual student data they can usually also obtain the individual demographic data. In such instances researchers employ more rigorous designs such as the matched student design (#2).

Designs 6-9 are based on comparisons of groups of students. Designs #6 and #7 follow same groups over time (e.g., 4th grade results in 2002 compared with 5th grade results in 2003). Designs #8 and #9 rely on comparing differing groups of students over time on the same grade level test (e.g., 4th grade reading students in 2002 compared with the next group of 4th grade students in 2003). The last 2 designs are cross-sectional in nature. More on these designs follows below.

Is the study longitudinal (i.e., does it follow students or groups of students over time)? Designs 1-9 are longitudinal. Longitudinal studies are important because they can illustrate impact of programs/reforms, and also they omit a lot of “noise” in the data from students entering or leaving the groups, since these students are typically dropped from longitudinal studies. The cross-sectional study designs noted in designs #10 and #11 can be misleading, since they provide snapshots of school performance that do not take into account the gain scores or growth in learning over time. Results from these studies can illustrate static performance of a group and they can illustrate the kind of students that are attracted to or are enrolled in the program. Cross-sectional studies with controls (#10) can illustrate how students are performing compared with demographically similar students in other schools. Cross-sectional studies are easier to do, because only one year of data is needed.

Does the study use student level data (designs #1-5)? Studies using student level data are stronger and more valid because the change in student performance over time is not influenced by students who have not received the treatment (e.g., students who take initial test but leave the school or students who arrive just before final posttest).

Many states do not have data files with individual student data or they legally are not allowed to release individual student data, so researchers must rely on group level data reported by grade and subject area. These are weaker designs, but when extensive controls are made, they can yield helpful data.

Controls for Comparison Groups

Key question to address:

What controls for comparison groups were used?

Below a list of typical controls is included. These student- or school-level variables can be used when matching students or schools. Also, they can be used when it is necessary to make statistical adjustments to account for background differences.

- | | |
|---|---|
| <input type="checkbox"/> Free and reduced lunch status (common indicator for poverty) | <input type="checkbox"/> Starting performance levels |
| <input type="checkbox"/> Ethnicity | <input type="checkbox"/> Urbanicity |
| <input type="checkbox"/> Special education or Limited English Proficiency | <input type="checkbox"/> Region |
| | <input type="checkbox"/> Parents' education level |
| | <input type="checkbox"/> Year of operation for school |

Because family income and poverty are strong determinants of education attainment and performance on standardized tests, the most important characteristics to match students should be free and reduced lunch status, which is a common indicator for poverty. In some instances, Title I status is also used as a indicator of poverty.

Ethnic background is also an important characteristic. Minority students often perform less well due to cultural differences not represented in standardized tests.

In terms of students' ability levels, three variables address the ability level or limitations of students. The first two are special education status and limited English proficiency status. In many studies, students in these groups are excluded from analysis because of the difficulty in making

suitable matches in comparison groups, and/or because the test used for these groups is different from the test administered to regular education students.

The starting performance of students is another variable focusing on ability. This is perhaps one of the most important controls that is commonly overlooked. Some school programs attract students from other schools that are unique in that they are predominately at risk of failure or are gifted. When a study takes into account the starting (pretest) performance level of students on a standardized test, it controls for relative ability. Students who are far below the mean are often able to demonstrate larger gains than students at or above mean performance levels.

Measures of Student Performance

Key questions to address:

What are the measures of student performance?

The measures of student performance are very important. Some measures are more precise for calculating gains for individual students over time. These include standard or raw scores on the test, national percentile rank, or normal curve equivalent.

Some measures are less precise such as quartile or decile scores, which will mask some change in scores if the students remain in the same quartile or decile. The crudest measures of student performance are cut-off scores such as those set according to state standards (e.g., student is at “basic level,” “meets standard,” or “exceeds standard”). The more precision a measure has, the more likely that researchers can capture and measure change.

Scope of the Study

Key questions to address:

What grades are included in the analysis?

What subject areas are covered?

What is the number and proportion of schools covered?

How many years of data are covered in the study?

Completeness and Quality of the Technical Report

Key questions to address:

Is there a technical report?

If so, does it explain the methods for the study?

Does the technical report spell out limitations of the study?

Does the technical report contain a full set of findings?

Based on information in the technical report, can the study be verified and replicated?

Reputation of Researcher and Host Organization

Key questions to address:

What is the reputation of the researcher?

Has she or he a vested interest in the outcomes of the study?

Has he or she come up with differing results from other studies?

What is source of the funding?

Does the researcher's host organization have a vested interest in outcomes?

The questions raised regarding the technical reports and the reputation of the researcher and host organization are critical issues to consider in terms of decisions to include or exclude studies for a synthesis of findings. However, because they are difficult to measure and apply in a weighting scheme one should avoid using them to weight and synthesize research.

Final Comments

The title of this paper emphasized that constructive uses for existing assessment data would be described and promoted in the contents of this work. The alternative approaches detailed in this paper promote the idea that we can work with less than desirable (albeit, best available) data. Other factors that have not been discussed but which also support the use of data such as those contained in the SSASD are costs and feasibility. The examples I have drawn on in this paper all represent large-scale or statewide evaluations with rather low costs for evaluating the impact of schools and reforms on student achievement. The most complicated study we conducted was with the matched student design in Delaware. The estimated costs for the student achievement component of this statewide evaluation of charter schools were around \$15,000. The estimated costs for the student achievement sections of the Pennsylvania and Edison studies were between \$20,000 and \$25,000. Compare these costs to the more than \$5 million that the US Department spent on a single study of charter school performance using a randomized experimental design.

Using existing data saves not only money but also considerable time. While it is necessary to wait several years for the findings from a randomized experiment, when existing data is used we can apply longitudinal designs by looking retroactively at existing data. When the findings from the federally-sponsored randomized study are released, it is certain to be one of the strongest and most defensible studies available. It is not, however, likely to provide a definitive answer regarding the performance of charter schools. Given the differences in charter schools within and between states, any definitive answer will come from a careful synthesis of the growing body of knowledge, most of which is based on readily available data generated from state assessment systems.

Appendix A

Sample Findings Using the Residual Gains Analysis

In this appendix, results from our evaluation of the Pennsylvania charter school initiative are included to illustrate the utility of the residual gains approach for measuring student achievement. The results yielded useful school level findings, as well as interesting aggregate findings for all charter schools and groups of charter schools.

Historically, Pennsylvania charter schools have scored much lower on the PSSA than noncharter schools. Knowing charter schools' achievement levels, however, tells us very little about their value as levers for improving student achievement. It is well known that student achievement scores reflect, in large measure, the "background" characteristics that students bring to the school. These include family income, race, special education status, urbanicity, and so on. In analyses conducted for this evaluation, these background factors typically accounted for close to three-fourths of the school-to-school variation in PSSA levels.

By themselves, the unadjusted achievement scores are more a measure of student characteristics than of school effectiveness. Indeed, the typical Pennsylvania charter school enrolls higher concentrations of disadvantaged students than other public schools. The challenge, therefore, is to determine what part of PSSA achievement scores reflect charter school effects, as opposed to the characteristics of the students who happen to enroll in them. In short, the question is how much educational value do charter schools add to their students?

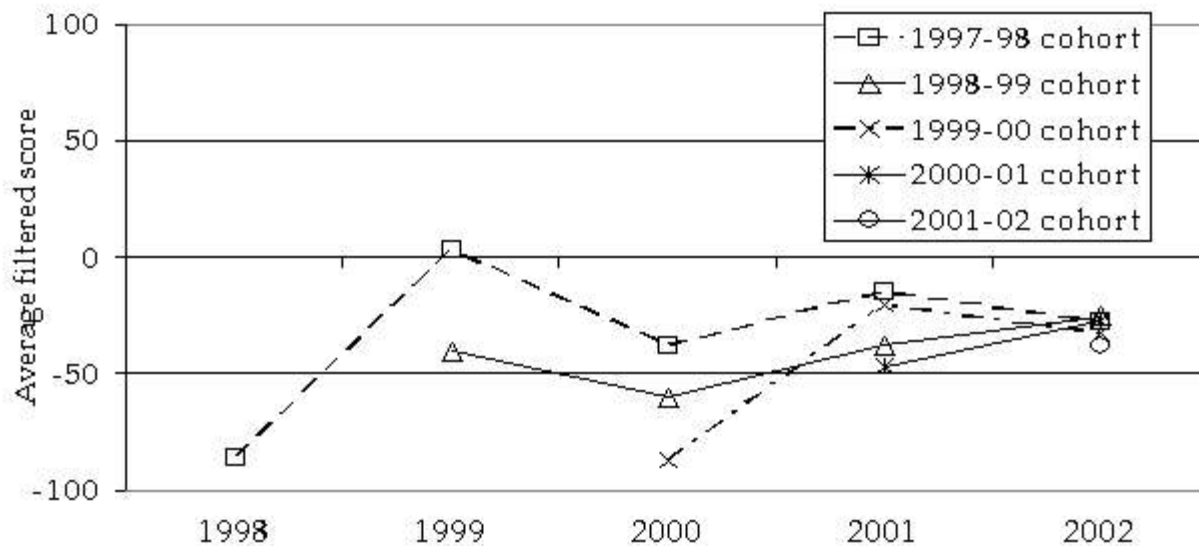


Figure A:1 Trends in Filtered Scores, by First Year of Operation

Note: A filtered score of zero indicates that the average charter school student scored exactly on par with the average student in demographically and geographically similar noncharter public schools. Positive (negative) filtered scores indicate that the average charter school student scored above (below) the average student in the comparison schools.

Using residual or filtered scores, the picture was brighter for charter schools than with the unfiltered PSSA scores. Before discussing changes in filtered scores over time, it is instructive to note that, averaged over time (and across grade levels and subject areas), the typical Pennsylvania charter school student scored just slightly lower (36 points) than the average student in his or her comparison group. Given that the PSSA scale ranges from approximately 1000 to 1600, a 36-point average deficit is a small one (5 percent of the scale range). In practical terms, whereas the average Pennsylvania charter school student scored some 140 points below the state average over the life of the initiative, he or she scored only slightly lower than the average student in demographically and geographically similar noncharter public schools.

Turning to changes in filtered scores over time, we found that 24 of the 42 schools (57 percent) with at least 2 years of PSSA data showed positive trends in filtered scores. Averaged over all schools, there was typically a 15 point gain in PSSA scores, after filtering out changes in student characteristics. Trends in filtered scores are illustrated in Figure A:1, which shows growth in average filtered scores for schools opened during the fall of 1997, 1998, 1999, 2000, and 2001. Figure A:2 provides a useful supplement to Figure A:1. Here are plotted scaled scores for both charter and comparison schools. Figure A:2 confirms the finding that the gap between charter and comparison schools is generally narrowing over time.

These trends are for aggregated scores at the school level that include all subject and grade level test results. Our technical report breaks out and

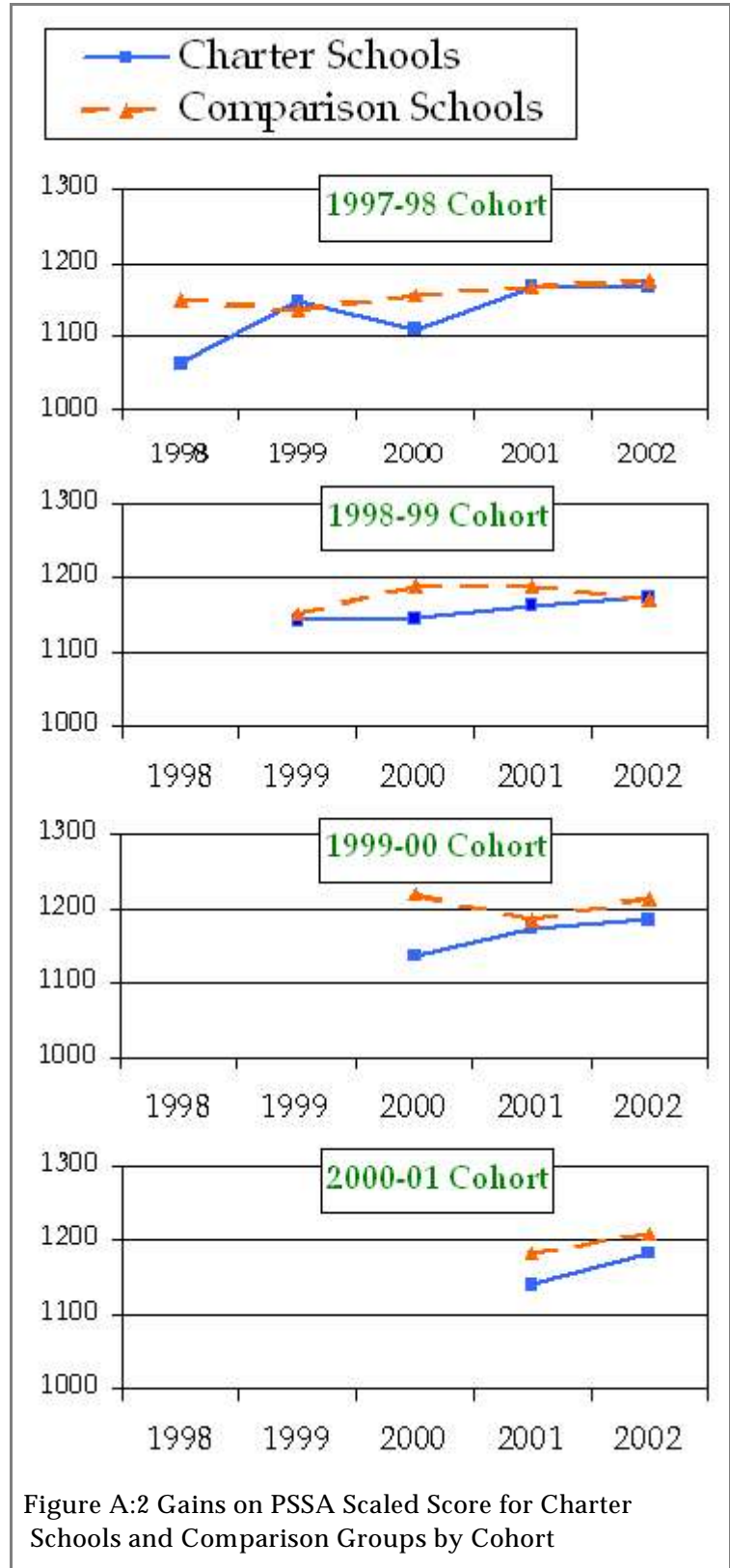


Figure A:2 Gains on PSSA Scaled Score for Charter Schools and Comparison Groups by Cohort

elaborates the differences by subject are and by grade. For the purpose of this report we can note that the charter schools did slightly better in reading than math, and the charter school students in grade 5 performed better than students in grades 8 and 11.

The cohort trends illustrated in the figures above—of course—mask considerable variance among charter schools. Figure A:3 illustrates the amount of variance in school performance.

Like other studies of achievement in charter schools, our analysis was subject to some important limitations. The most important of these is that the PSSA is not well suited to tracking student gains over time. Indeed, instead of tracking a single cohort of students over time, analysis is restricted to the comparison of this year’s fifth, eighth, or eleventh graders with different groups of fifth, eighth, and eleventh graders in subsequent years. Thus, evaluators must find some

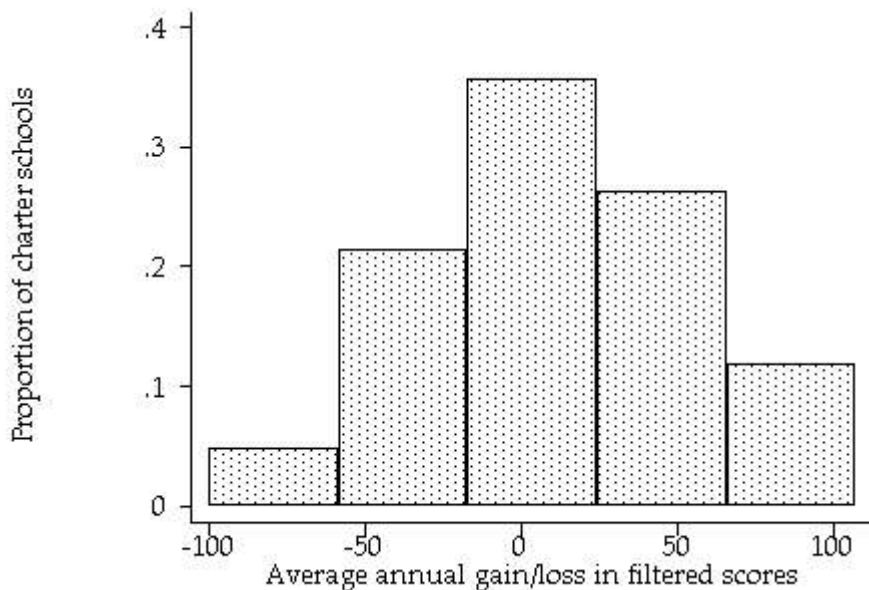


Figure A:3 Variation in Annual Growth Rates Across Charter Schools

way to distinguish score changes that reflect school effectiveness from those that reflect year-to-year changes in student composition. To this end, we have developed a statistical “filtering” methodology that subtracts the influence of student background factors. While not foolproof, this method represents a substantial improvement over examination of unadjusted PSSA scores.

Another important limitation of these findings is that, short of a randomized experiment, one cannot be absolutely sure that the charter-noncharter differences have been caused by the charter school law. To be sure, the system of demographic and geographic controls used to derive the filtered scores rules out a large number of rival explanations for the differences. However, it is impossible to rule out the possibility that the charter-noncharter differences are due to unmeasurable differences in parental motivation, social capital, and other intangible factors. Once again, our methods, while not foolproof, represent a considerable improvement from the examination of unadjusted test scores.

Even stronger analyses of the charter school achievement impact will be possible should Pennsylvania move to a system that facilitates the tracking of individual achievement gains over time. Also, the passage of more time will provide longer series of data against which to estimate more certain growth trends. In the meantime, policymakers must evaluate the initiative’s effectiveness with the data at hand. The findings in this study were designed to provide sound data to inform—though not fully justify—decisions about the Pennsylvania charter school initiative.

Appendix B

Sample Findings Using Odds Ratio Analysis

The primary purpose of this evaluation was to develop a composite understanding of the effect of attending an Edison school on students' achievement. This involved cross-state analyses, which resulted in us having to use similar but less desirable comparison groups. Comparison groups were defined in such a way that the data sources would represent the same quality of information and thus have similar meaning from state to state. Also, we had to use a common measure of performance (cut scores) which could be standardized across the states, even when some states have better and more sensitive measures of school performance on standardized tests.

Construction of the comparison groups. We were interested in examining the number/proportion of Mid-Michigan PSA students who met state standards ("passing") or conversely the number/proportion of students who did not meet state standards ("failing") on the Michigan Education Assessment Program (MEAP). In order to determine relative impact, we needed to define a suitable comparison group. In the grade 4 and 5 analyses, our first comparison was with the Lansing School District in which the Edison school resided. The Edison school and the local district had student with similar demographic backgrounds.

State performance constituted our second comparison group. While the state demographics vary from Mid-Michigan Public School Academy and the Lansing Public School District, we believe that comparisons with state averages can yield further information regarding the relative gains of the Edison school. Also, since Edison claims that advances in other district schools is—in part—due to its presence, we used the state as a more distant point of comparison that cannot be easily influenced by the presence of Edison schools.

General procedure. Utilizing published data from the state of Michigan, we made yearly comparisons (consecutive cohorts) at grade 4 and grade 5 from 1997 through 1999 and from 1998 to 1999 in grades 7 and 8 for each subject component of the MEAP test administered at those specific grade levels.

Percentage data (students in each scoring category) were converted to raw frequencies and the multilevel scoring was collapsed into a dichotomy (pass, fail), thereby producing 2x2 contingency tables. To ensure independence of the rows in these tables, the raw frequencies for each scoring category of the MEAP in the district and state comparisons were down-weighted by subtracting the number of students in that category from Mid-Michigan.

One of the many possible statistics that can be derived from a 2x2 contingency table is the odds ratio statistic and corresponding $1-\alpha$ confidence interval. From a classical epidemiological perspective, the students in the "Edison" school can be thought of as the "exposed" group—that is, exposed to the "Edison-effect"—and students in the comparison group as the unexposed group. From this perspective, each yearly comparison is a "new" cohort; and, measured over a period of years, there are consecutive class cohorts. There is a minimal possibility for cohort contamination if a number of students in one group are not promoted to the next grade level. However, this represented a very small number of possible cases and has minimal impact on the validity of these analyses. We calculated and charted OR for each of the 2x2 tables constructed from the chi-square analyses presented above. The findings for grades 4 and 5 are included below and help illustrate the utility of this strategy.

Odds ratio findings for grade 4. In grade 4 reading (see Exhibit B-1), the OR for Mid-Michigan showed relative stability in magnitude against the district, revealing a slight rise in OR in 1999. The Breslow-Day statistic for testing the hypothesis of homogeneity of OR over the three years was not statistically significant, indicating there was no real (statistically significant) change in the OR over the three years. The common OR for the three years is 1.672 and the 95 percent CI is from 1.284 to 2.176. Since the CI does not include 1.00, there was a statistically significant increase in odds for a Mid-Michigan student to fail the grade 4 MEAP reading test relative to students in the district. That is, the Mid-Michigan students were about 1.6 times more likely to fail this test relative to students in the district. A similar pattern in OR is observed relative to the state except that the magnitude is larger, indicating that the Mid-Michigan students were even more likely to fail the reading MEAP test relative to the state. The Breslow-Day statistic for testing the hypothesis of homogeneity of OR over the three years was not statistically significant, indicating there was no real (statistically significant) change in the OR over the three years. The common OR is 3.249 and the 95 percent CI is from 2.549 to 4.142, indicating that Mid-Michigan students were about 3¼ times more likely to fail this test.

The grade 4 math component of the MEAP presented a similar picture. All the CI around the individual ORs exclude 1.00 and thus are considered statistically significant. Likewise, the Breslow-Day statistic for testing the hypothesis of homogeneity of OR over the three years was not statistically significant, indicating there was no real (statistically significant) change in the OR. The common OR relative to the district is 2.242 and the 95 percent CI is from 1.748 to 2.876. For the state comparison the common OR is 4.553 and the 95 percent CI is from 3.639 to 5.698. The Mid-Michigan students were about 2¼ times more likely than district students to fail the math test over the three years and about 4½ times more likely to fail relative to the rest of the state.

Odds ratio findings for grade 5. In grade 5 science (see Exhibits B-2), the OR for Mid-Michigan showed relative stability in magnitude against the district. The Breslow-Day statistic for testing the hypothesis of homogeneity of OR over the three years was not statistically significant, indicating there was no real (statistically significant) change in the OR. The common OR for the three year period is 2.411 and the 95 percent CI is from 1.709 to 3.401. Thus, there was a statistically significant increase in odds for a Mid-Michigan student to fail the Grade 5 MEAP science test relative to students in the district in this three year period. Mid-Michigan students were almost 2.5 times more likely to fail this test relative to students in the district. A similar pattern in OR is observed relative to the state except that the magnitude is larger, indicating that the Mid-Michigan students were even more likely to fail the science MEAP test relative to the state. The Breslow-Day statistic for testing the hypothesis of homogeneity of OR over the three years was not statistically significant, indicating there was no real (statistically significant) change in the OR over the three years. The common OR is 4.228 and the 95 percent CI is from 3.078 to 5.809, indicating that Mid-Michigan students were about 4¼ times more likely to fail this test as compared with the students in the state.

The grade 5 writing component of the MEAP presented a slightly different picture. In the district comparison the Breslow-Day statistic for testing the hypothesis of homogeneity of OR over the three years was statistically significant, indicating that the OR needs to be examined each year due to its changing value. As can be seen in Exhibit 11:4, the OR in 1999 took a substantial jump. This rising pattern is also apparent in the state analysis, although it did not reach statistical significance. The Breslow-Day statistic for testing the hypothesis of homogeneity of OR over the three years was not statistically significant, and the common OR for the three year period was 2.436 and the 95 percent CI was from 1.930 to 3.076.

EXHIBIT B:1 Results of the Odds Ratio Analyses, Grade 4

Mid Michigan Grade 4 MEAP Reading vs Lansing

Year	UB	LB	OR
1997	2.221	0.877	1.396
1998	2.289	0.948	1.473
1999	3.680	1.433	2.296

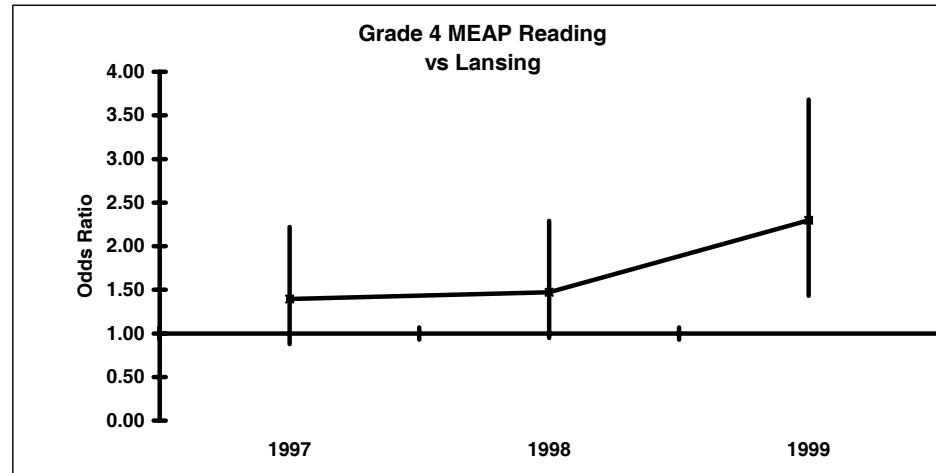
Breslow-Day for Homogeneity of Odd Ratio

Chi-Sq (2, N=3,845) = 2.564, p = .277

OR = 1.672

LB = 1.284

UB = 2.176



Mid Michigan Grade 4 MEAP Math vs Lansing

Year	UB	LB	OR
1997	3.163	1.302	2.029
1998	3.482	1.501	2.286
1999	3.733	1.584	2.432

Breslow-Day for Homogeneity of Odd Ratio

Chi-Sq (2, N=3,856) = 0.327, p = .849

OR = 2.242

LB = 1.748

UB = 2.876

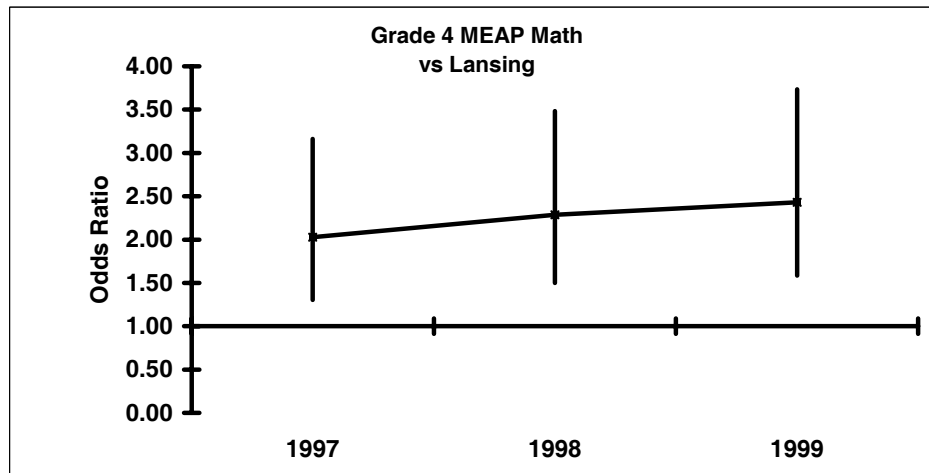


EXHIBIT B:2 Results of the Odds Ratio Analyses, Grade 5

Mid Michigan Grade 5 MEAP Science vs Lansing

Year	UB	LB	OR
1997	4.062	1.085	2.099
1998	3.890	1.287	2.238
1999	5.249	1.611	2.908

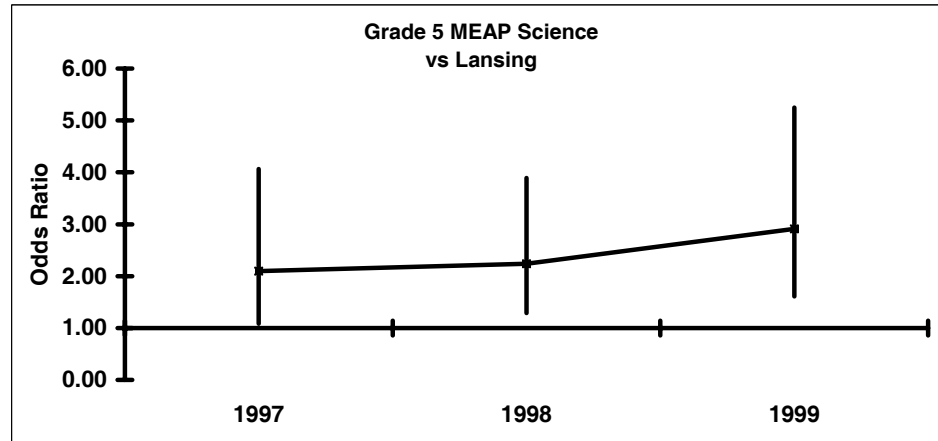
Breslow-Day for Homogeneity of Odd Ratio

Chi-Sq (2, N=3,651) = 0.587, p = .746

OR = 2.411

LB = 1.709

UB = 3.401

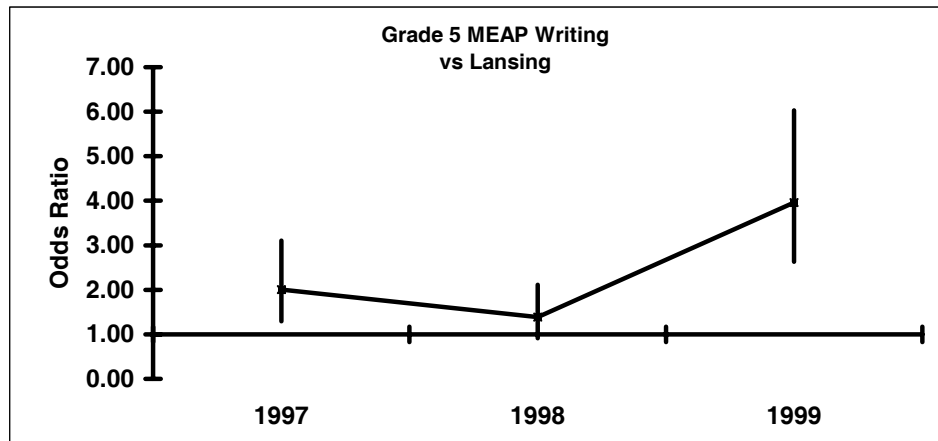


Mid Michigan Grade 5 MEAP Writing vs Lansing

Year	UB	LB	OR
1997	3.105	1.291	2.002
1998	2.110	0.914	1.389
1999	6.030	2.629	3.960

Breslow-Day for Homogeneity of Odd Ratio

Chi-Sq (2, N=3,594) = 11.998, p = .002



Appendix C

Sampled Findings from a Study Using a Matched Student Design

In this appendix, some of our findings from our use of a matched student design are included to provide greater insight into the utility of this design.

Table C:1 contains the results from our analysis that incorporated all charter school students. There are two panels and two subjects (i.e., reading and mathematics) for each grade, which means that there were four analyses at each grade level. We did not aggregate the results by grade or subject. Instead, we have reported the results from each analysis separately. In our description and discussion of the findings, we drew conclusions by grade and subject.

The results were reported by grade and subject area and included both scaled score results on the DSTP and the normal curve equivalent (NCE) scores on the SAT-9. [A number of items from the SAT-9 are incorporated in the DSTP test so that equivalent scores can be calculated for the SAT-9.] Therefore, while the scaled score results reflect total scores on the DSTP, the NCEs reflect performance on a subset of questions. This can explain differences in relative performance levels that exist between the two sets of scores.

The *covariate mean* (see Table C:1) is the mean score for all students in the group in the prior DSTP assessment. Therefore, the covariate mean for students in grade 5 would be their scores two years earlier in grade 3. The *adjusted mean* is the focus of the ANCOVA analysis, the second DSTP assessment. This is not the observed mean score (weighted mean) for the group; rather, it is a mean score adjusted for students' performance on the prior assessment. The ANCOVA provided two statistical tests: one for the covariate (slope of the relationship between the prior assessment and the target assessment is non zero) and one for the adjusted means (the hypothesis of interest). If the covariate is found to be statistically significant, then the ANCOVA will allow a more powerful test of the adjusted means, which is the second hypothesis considered in the model. Evaluation of the covariate should always be considered and in all analyses was found to be statistically significant. Thus, the use of the ANCOVA was justified in that there was a statistically significant relationship between the prior DSTP assessment and the target DSTP assessment. In Table C:1 the F-value and associated p-value reported correspond to the hypothesis of no difference between the adjusted (target) DSTP means (charter vs non charter). If the F-value is large and the corresponding p-value small it is common practice to reject the hypothesis of no difference in favor of the alternative hypothesis: there exists a difference in the adjusted DSTP means between charter and non charter schools.

The ANCOVA carries two important statistical assumptions which should be carefully examined for valid interpretation. The first is the assumption of homogeneity of variance and the second is the homogeneity of regression slopes. Of the 24 analyses presented in Table C:1, in only one analysis the assumption of equal slopes was violated and in four analyses the equal variance assumption was violated.

The results in Table C:1 indicate that the charter school students performed better than matched traditional public school students in the upper grades. There were small differences between the charter school students and comparison students between grades 3 and 5. Only two differences were statistically significant; one of these differences favored traditional public schools, and the other difference favored charter schools. At grade 8, two of the four comparisons proved to have large differences that were

Table C:1 Performance on DSTP for Charter School Students and Comparison Students by Grade and Subject Area

Grade and Subject Area	Scaled Score on the DSTP				Normal Curve Equivalent on the SAT-9			
	Covariate Mean	Adjusted Mean	F-value	P-value	Covariate Mean	Adjusted Mean	F-value	P-value
Grade 5 Reading, Panel A								
Charter school	442.3	483.2	0.02	0.8853	58.6	57.8	5.84	0.0158
Control group	446.8	483.4			61.2	55.8		
Grade 5 Reading, Panel B								
Charter school	435.9	482.5	0.17	0.6775	57.2	56.0	0.39	0.5309
Control group	439.5	481.8			58.3	55.3		
Grade 5 Math, Panel A								
Charter school	435.2	471.2	8.21	<u>0.0043</u>	61.1	63.2	2.28	0.1312
Control group	435.3	475.5			62.9	61.7		
Grade 5 Math, Panel B								
Charter school	428.9	466.8	0.20	0.6530	59.4	59.1	0.00	0.9540
Control group	431.9	467.5			61.0	59.0		
Grade 8 Reading, Panel C								
Charter school	484.7	532.8	1.81	0.1787	58.5	64.3	6.61	0.0104*
Control group	479.9	530.5			58.9	61.4		
Grade 8 Reading, Panel D								
Charter school	486.1	531.6	1.41	0.2348	60.3	62.2	0.09	0.7697
Control group	478.0	529.5			57.3	61.9		
Grade 8 Math, Panel C								
Charter school	474.6	513.0	7.56	0.0061*	64.2	64.3	5.86	0.0157*
Control group	468.5	508.2			60.1	61.5		
Grade 8 Math, Panel D								
Charter school	477.0	509.0	1.36	0.2434	63.3	61.5	2.05	0.1527
Control group	469.1	511.2			61.3	59.8		
Grade 10 Reading, Panel E								
Charter school	550.2	544.5	20.30	>.0001*	72.3	62.3	34.42	>.0001*
Control group	532.6	534.5			63.8	54.7		
Grade 10 Reading, Panel F								
Charter school	550.8	540.0	3.29	0.0704	74.3	62.3	17.68	>.0001*
Control group	528.3	535.6			64.4	56.1		
Grade 10 Math, Panel E								
Charter school	539.5	564.1	7.75	0.0056*	74.6	69.4	1.76	0.1853
Control group	510.1	556.2			62.2	67.3		
Grade 10 Math, Panel F								
Charter school	534.7	563.1	22.35	>.0001*	75.2	68.8	8.54	0.0037
Control group	505.7	550.2			60.0	64.0		

Notes. Comparison group is matched on gender, ethnicity, FRL, and Title I status.

Differences between the charter school students and comparison students are statistically significant when the P-value is less than 0.05; these scores are highlighted in **bold**. When P-values are underlined and bolded, this refers to an advantage to the noncharter school students.

P-values with an asterisk “*” refer to differences that remained statistically significant at least 80 percent of the time with repeated randomly selected comparison groups.

statistically significant. These differences were for Panel C (not Panel D) and all of these differences favored charter schools.

The largest differences between charter school students and matched students in traditional public schools were at grade 10. Here three of the four comparisons showed that the differences were statistically significant, and all these differences favored charter school students (Panel F Reading had significant differences favoring charter schools on the SAT-9 items, but not on the overall DSTP). In other words, the charter school students included in the panels were gaining more on the DSTP between grade 8 and grade 10 than traditional public school students. The differences that were significant at grades 8 and 10 typically were larger and remained statistically significant even after we generated additional randomly selected comparison groups. One serious limitation to keep in mind here is that many students in the grade 8 to grade 10 panels did not actually enter a charter school until grade 9. Also many students were dropped from this panel because they did not have a grade 8 DSTP score. This is likely because they were enrolled in private schools or were coming from out of state.

Where differences were especially large and significant on the DSTP scaled score, the difference on the NCE for the SAT-9 subset of items was also statistically significant. When the differences were small but still statistically significant, it often happened that only the scaled score or only the NCE score proved to be statistically significant.

The panels that included more recent years of data (i.e., Panels A, C, and E which ended in 2004) showed more differences that favored charter schools than the more earlier panels (Panels B, D, and F which ended in 2003). This provides some tentative evidence that charter schools are improving over time. However, this may also be explained by the fact that the more recent panels include more schools, some of which have fewer years of operation. Over time, the Department of Education has raised the bar in terms of new applicants which may explain why more recently established charter schools help lift the performance of Panels A, C, and E). The full technical report uses the same approach to examine the performance of individual charter schools.

Creaming the best or serving the neediest? The data in Table C:1 illustrate important information about the types of students attracted to charter schools. While many charter schools establish curricular profiles and marketing materials that make them most attractive to students failing in traditional public schools, some charter schools also have profiles and marketing practices that help them attract high performing students. The covariate means in the table represent the pretest scores of the students that are matched by race, free and reduced lunch status, English Language Proficiency status, and Title I status. When the covariate mean for the charter school group and control group is similar, this means that the charter school has students who are performing similarly to their demographically matched peers. When the charter school group has a higher covariate mean than the control group, this indicates that the enrolled charter school students already have higher performance levels at the time of pretest.

A comparison of the covariate means at grade 4 illustrates that the charter school students and demographically similar students in the control group have similar pretest performance levels. At grade 8, the charter schools are clearly attracting and enrolling higher performing students. This difference is further exacerbated in grade 10, where the charter school students have substantially higher pretest scores than their demographically similar peers. These comparisons suggest that while the charter schools on the whole are not “creaming” or attracting the best performing students in lower elementary grades, they clearly are doing so in the lower and upper secondary levels.

The data in Table C:1 are aggregated across all the schools, which masks large differences between the schools, both in terms of the students they enroll and in terms of the growth in test scores they can affect. As it turned out, the results varied considerably by school with a few of the larger schools performing quite well and a number of the smaller schools with mixed or negative findings.

References

- AFT. (Nelson, F. H.; Rosenberg, B., & Van Meter, N.) (2004.) *Charter school achievement on the 2003 National Assessment of Educational Progress*. Washington, D.C.: American Federation of Teachers.
- Carnoy, M., Jacobsen, R. Mishel, L., & Rothstein, R. (2005). The charter school dust-up. Examining evidence on enrollment and achievement. Washington, D.C.: Economic Policy Institute.
- Gill, B. P., Timpore P.M., Ross, K.E., & Brewer, D.J. (2001). *Rhetoric versus reality: What we know and what we need to know about vouchers and charter schools*. Santa Monica, CA: RAND.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research*, 5, 3-8.
- Hassel, B. C. (2005). Charter school achievement: What we know. Paper prepared for the Charter Schools Leadership Council. <http://www.charterschoolleadershipcouncil.org/PDF/paperupdate.pdf>
- Hill, P. T. (2006). Assessing achievement in charter schools. In R. Lake & P.T. Hill (eds.) *Hopes, fears, & reality: A balanced looks at American charter schools in 2005*. Seattle, WA: Center on Reinventing Public Education, University of Washington.
- Hoxby, C. M. (2004). "A straightforward comparison of charter schools and regular public schools in the United States." <http://post.economics.harvard.edu/faculty/hoxby/papers/hoxbyallcharters.pdf>.
- Hoxby, C.M. & Rockoff, J. E. (2004, May). The impact of charter schools on student achievement." <http://post.economics.harvard.edu/faculty/hoxby/papers/hoxbyrockoffcharters.pdf>.
- McNeil, D. (1996). *Epidemiological research methods*. New York: Wiley.
- Miron, G. (2004). Evaluation of the Delaware charter school reform: Year 1 report. Dover, DE: The Delaware Department of Education. http://www.wmich.edu/evalctr/charter/de_cs-eval_year1_report.pdf.
- Miron, G. (2005.) Evaluating the performance of charter schools in Connecticut. A report commissioned by the Connecticut Coalition for Achievement Now (ConnCAN). http://www.wmich.edu/evalctr/charter/ct_charter_school_performance_2005.pdf
- Miron, G., & Applegate, B. (2000). An evaluation of student achievement in Edison schools opened in 1995 and 1996. Kalamazoo, MI: The Evaluation Center, Western Michigan University. [Online]. <http://www.wmich.edu/evalctr/edison/edison.html>.
- Miron, G., & Horn, J. (2002). Evaluation of Connecticut charter schools and the charter school initiative: Final report. Kalamazoo, MI: The Evaluation Center, Western Michigan University. [Online]. <http://www.wmich.edu/evalctr/charter/ctcharter.html>.
- Miron, G. & Nelson, C. (2001). Student academic achievement in charter schools: What we know and why we know so little. Occasional Paper No. 41. New York: National Center for the Study of Privatization in Education. Teachers College, Columbia University. http://www.ncspe.org/publications_files/590_OP41.pdf.
- Miron, G., & Nelson, C. (2002). *What's public about charter schools: Lessons learned about school choice and accountability*. Thousand Oaks, CA: Corwin Press.
- Miron, G. & Nelson, C. (2004). "Charter schools and academic achievement." In K. Bulkley & P. Wohlstetter (eds.). *Cutting loose: Autonomy and education in charter schools*. NY: Teachers College Press.
- Miron, G., Nelson, C., & Risley, J. (2002). Strengthening Pennsylvania's charter school reform: Findings from the statewide evaluation and discussion of relevant policy issues. Kalamazoo, MI: The Evaluation Center, Western Michigan University. http://www.wmich.edu/evalctr/charter/pa_5year/.
- NCES. (2005). The nation's report card. America's charter schools: Results from the NAEP 2003 pilot study. Washington, D.C.: National Center for Education Statistics.
- Slavin, R.E. (1986) Best-evidence synthesis: An alternative to meta-analytic and traditional reviews, *Educational Researcher*, 15 (9), pp 5-11.
- Solmon, L., Paark, K., & Garcia, D. (2001). *Does charter school attendance improve test scores? The Arizona results*. Phoenix, AZ: Goldwater Institute.