Comparing Achievement between K–8 and Middle Schools: A Large-Scale Empirical Study

VAUGHAN BYRNES and ALLEN RUBY
Johns Hopkins University

This study compares middle schools to K–8 schools, as well as to newly formed K–8 schools that are part of a K–8 conversion policy. The outcome is student achievement, and our sample includes 40,883 eighth-grade students from 95 schools across five cohorts. The analysis uses multilevel modeling to account for student, cohort, and school-level variation, and it includes statistical controls for both population demographics and school characteristics. The results find that older K–8 schools perform significantly better than middle schools, and this advantage is explained by differing student and teacher populations, average grade size, and school transition. Newer K–8 schools did not enjoy the same advantage despite having smaller grades and lower transition rates, due to their more disadvantaged populations.

This article reports on a natural experiment in the Philadelphia City School District. Using longitudinal data and multilevel modeling, we compare students who attended middle schools to students from K–8 schools in order to determine if the different school structures had any effects upon students' mathematics and reading achievement, and if so to assess the particular causes of any such differences.

The reason for such a study is that middle-grades education in the United States has struggled in terms of academic achievement, especially among urban public schools and those schools serving high-minority and high-poverty students (Beaton et al. 1996; Schmidt et al. 1999). As the gap between the more advantaged and disadvantaged students within the United States has widened, along with the gap between the United States itself and other developed nations, many large-scale and high-resource reform efforts have been undertaken over the last decade with the direct aim of improving student achievement in these schools (Burrill 1998). One of the more popular reforms currently

Electronically published August 16, 2007
American Journal of Education 114 (November 2007)

0195-6744/2007/11401-0004$10.00

sweeping across the educational landscape is a policy of converting middle schools into K–8 schools, with the belief that the latter are more effective at nurturing student achievement. However, while the policy of K–8 conversion has quickly gained steam, the research on the subject matter has to this point lacked the large and rigorous statistical research needed to provide scientific evidence for supporting such a policy.

In this study, we compare the K–8 and middle schools of Philadelphia, one of the largest public school districts in the United States and one that serves a predominantly high-minority and high-poverty student population. It is also a school district that is currently implementing a K–8 conversion policy. By analyzing this quasi-experimental setting, using an appropriate method of statistical analysis, combined with a large sample taken over a wide time span, and by incorporating the most theoretically relevant statistical controls, we hope to provide empirical evidence either confirming or disproving the hypothesis that K–8 schools perform better in terms of student achievement and, if confirming it, to provide theoretically strong and empirically supported causal explanations for it. We hope that the results, aside from adding to the wider scientific literature, will be of use to policy makers, districts, and schools who are considering a policy of K–8 conversion, better informing them on what results they can expect from such a reform and under what circumstances results may vary.

A Return to the K–8 School

Of the many reforms to become popular in middle-grades education, the converting of middle schools into K–8 schools must be one of the more remarkable. What makes this reform so interesting is that where most reforms attempt some new innovation, the conversion to the K–8 model represents in some way a return to the old.

VAUGHAN BYRNES is a researcher at the Center for Social Organization of Schools at Johns Hopkins University. He received his degree in social research methods and statistics from the London School of Economics. ALLEN RUBY is an associate research scientist at the Center for Social Organization of Schools (CSOS) of the Johns Hopkins University. His work has focused on the implementation of curriculum, teacher professional development, the reform of urban middle schools, and the evaluation of the impact of school reform on student outcomes. The research reported here was supported by the Research on Learning and Education (ROLE) Program at the National Science Foundation, grant no. 0411796; a CSR Quality Initiative grant from the Institute of Education Sciences, U.S. Department of Education; and an Interagency Education Research Initiative grant (r305w020003).

It was the K–8 school that predominated middle-grades education in the United States at the end of the nineteenth century and through the first four decades of the twentieth, only to be supplanted by the junior high model (grades 7–9) that began in the 1910s and became the predominant school structure by the 1960s. The junior high was then itself replaced by the middle school model, which rose in the 1960s and 1970s to become the dominant school structure of the 1990s (Herman 2004; Mizell 2005; Paglin and Fager 1997). Now, however, the K–8 structure is once again the popular choice and middle school conversions are quickly sweeping the nation.

Already, reforms have begun in states such as Massachusetts, Pennsylvania, Ohio, Tennessee, Oklahoma, Maryland, and New York, including the large urban districts of Cincinnati and Cleveland, Philadelphia, and Baltimore, while other districts looking to convert their middle schools into K–8s include at least eight other states across the nation (Hough 2005; Pardini 2002; Reising 2002). In some ways the K–8 model has also never left, as it has remained a popular choice among private and parochial schools, as well as in several European countries (Herman 2004).

Why Go Back?

When the middle school model was first established, it was with the notion that by isolating those middle grade years, the schools would be perfectly suited to handling both the academic and emotional needs of those early adolescents to which they would cater.

On the one hand, isolating early adolescents was to allow the schools a chance to focus on the behavioral needs specific to 10–13-year-olds (Coladarci and Hancock 2002; Yakimowski and Connolly 2001). On the other, such middle schools would be able to engage in a set of "best practices"—pedagogical strategies, instructional strategies, small learning communities (or schools within schools), professional development for teachers, team teaching (or semidepartmentalization), and mixed-level classrooms—that would allow them unique advantages in addressing the academic achievement of middle-grades students (Epstein and MacIver 1990; Hough 2005; Lee and Smith 1993; Midgley 1993; Offenberg 2001). However, over the last decade or so, research has been put forth that would suggest the opposite to be true, that middle-grades students attending K–8 schools show distinct advantages over middle school students in both academic and nonacademic areas.

First and foremost, some research has shown that students at K–8 schools have higher levels of academic achievement, both in mathematics and reading (Coladarci and Hancock 2002; Offenberg 2001; Yakimowski and Connolly 2001). As academic achievement in terms of performance on standardized

tests is the fundamental method of evaluation for both districts and states in assessing school performance and reform efforts, such evidence has been the most common source of validation for school conversion efforts. In addition, however, students attending K–8 schools have also been found to have higher rates of attendance (Coladarci and Hancock 2002; Pardini 2002) and better performance in terms of emotional and social outcomes such as self-esteem, leadership, and attitudes toward school (Simmons and Blyth 1987; Weiss and Kipnes 2006). These social engagement and attitudinal outcomes are extremely important, not only as outcomes themselves but because they in turn then have effects on student achievement.

Student outcomes, though, are not the only validation for conversion efforts. Parents often praise the greater sense of community that they feel exists in K–8 schools, and several studies have noted the stronger relationships that seem to exist between students, between teachers, between students and teachers, and between parents and teachers in K–8 schools (Herman 2004; Pardini 2002; Offenberg 2001; Yakimowski and Connolly 2001). That K–8 schools are often closer to home in terms of travel is also an aspect that parents appreciate, and that the schools are then even more of a local neighborhood school adds to their greater sense of community (Herman 2004; Mizell 2005). In addition, parents also like that the longer grade span allows for families with several children to have siblings in the same school for longer periods of time (Pardini 2002).

Districts, too, have other reasons for preferring the K–8 structure. One is that when making the transition from an elementary school to a middle school, many students leave the district entirely in what is considered to be a flight from failing urban public school systems, a trend that some districts hope to stem with K–8 conversions (Herman 2004; Yakimowski and Connolly 2001). Another reason for districts to convert is that K–8 schools are often more cost efficient in terms of building and property maintenance (one building vs. two): an important factor when fighting stretched budgets (Herman 2004).

Thus, students, parents, and districts might all stand to benefit from K–8 conversions if the above-stated advantages are in fact true. Yet it is not enough to simply believe that K–8 schools are better, and while their advantages seem diverse, the explanations for them must also be made clear if policies based upon them are to be widely applied.

What Makes the K–8 Schools Better?

According to current theories, the K–8 advantage centers around two sets of main causal factors. The first set pertains to schools' population demographics and are external to the type of grade structure they might have, while the

second set of factors are directly related to a school's choice of grade structure. These factors are differences between the student populations of K–8 and middle schools, differences between the teacher populations common to the two school structures, the extra transition to a new school that middle school students must make at the end of elementary school, and differences in the average size of K–8 schools versus middle schools.

Student Demographics

The first, that middle schools in general serve student populations with higher rates of poverty and larger proportions of minority students, is one of the fundamental reasons suggested by prior research as to why the two school structures might show different levels of aggregate achievement (Balfanz 2002; Offenberg 2001; Yakimowski and Connolly 2001). Poor students from minority backgrounds are likely to have a harder time both in and out of school, due to language barriers, lack of resources, less stable homes, and the turbulence of disadvantaged neighborhoods, leading to poorer performance in school. Such demographic factors likely affect attendance rates and other social outcomes, as well as academics. If student demographics are the main reason for the different academic performances of the two school types, then converting middle schools into K–8s may not lead to a significant improvement in student achievement if the student population remains unchanged.

Teacher Population

Similarly, teacher characteristics such as years of experience, levels of certification, retention rates, and student-teacher ratios in middle schools are also thought to be different from those in K–8 schools, contributing to the schools' diverging performances on student achievement and social outcomes (Paglin and Fager 1997; Simmons and Blyth 1987). As most teachers are trained to teach at either the elementary or high school level, many middle school staffs are then faced with lower rates of retention, less experience, and lower rates of certification, as those with seniority transfer out to the elementary and high schools for which they are certified (Jackson and Davis 2000; McEwin and Dickinson 1996; McEwin et al. 1996; National Forum to Accelerate Middle Grades Reform 2002). The lack of middle-level trained and certified teachers may also have prevented the middle school model from being properly implemented in the first place, given the uniqueness of adolescent students and the particular set of teaching practices that were recommended for teachers of middle grades, which may require more specific training.

School Transition

Another factor that might affect both academic and social differences between school structures, but one that is intrinsic to a school's grade structure, is the extra transition to a new school that middle school students must make. First, it seems clear that when moving to a new school students must adjust to a new environment where they do not know, and in turn are not known by, most other students and staff, a change that can be harmful for student performance and engagement, especially among minority students (Coladarci and Hancock 2002; Herman 2004; Simmons and Blyth 1987; Simmons et al. 1991).

An offshoot of this transition is that when middle school students make the switch from their elementary school, they also enter as the youngest children in their new building. Some past research has found that K–8 students may benefit from spending the middle grades as the older children in their school building, and that being the "top dog" might lead to greater feelings of confidence, maturity, and leadership (Coladarci and Hancock 2002; Herman 2004; Simmons and Blyth 1987; Yakimowski and Connolly 2001). Thus the change between schools can have both direct and indirect effects on students' academic achievement and other social and engagement outcomes.

School Size

Most important, however, both practically and theoretically, might be the size of the school. Previous studies have linked size to virtually all of the K–8 advantages, with the larger size of middle schools being detrimental to the student outcomes of academic achievement, attendance, and social engagement (Coladarci and Hancock 2002; Eccles et al. 1991; Lee and Smith 1993; Offenberg 2001; Simmons and Blyth 1987; Weiss and Kipnes 2006). Fundamentally, students in smaller schools are subjected to less anomie and receive more personal attention from their teachers. In a small school, it is easier for students to get to know and develop mutual respect for each other, while it is also easier for teachers to be familiar with students who aren't their own and know when they shouldn't be wandering the hallways. Thus, smaller size is also a probable cause of the stronger sense of community visible in K–8 schools, as smaller size allows students, teachers, and parents to foster closer and stronger relationships (Offenberg 2001; Paglin and Fager 1997).

With a touch of irony, smaller size may also enable K–8 schools to more effectively implement the very set of "best practices" that were originally thought to be an advantage of middle schools, and the greater use of these practices may also be a reason why K–8 schools tend to perform better. Several

studies have found these activities to be more common in K–8 schools than in middle schools, and smaller populations may be more conducive to creating personal learning communities, having teachers coordinate for team teaching, and establishing mixed-level classrooms (Coladarci and Hancock 2002; Eccles and Midgley 1989; Hough 2005; Offenberg 2001; Yakimowski and Connolly 2001). If the advantage of K–8 schools over middle schools is due more to these intrinsic factors of grade structure such as school continuity and smaller size, as opposed to external population demographics, then the policy of converting middle schools into K–8 schools is one that should be of benefit.

This Study

While the existing research has been clear on what the advantages of K–8 schools over middle schools are and for what reasons they may exist, the actual amount of research that has been done is quite small considering how widely the policy of K–8 conversion is being adopted across the United States. Of the research that has been completed, even less has employed rigorous statistical and thorough empirical techniques, with most of it based upon case studies and descriptive or anecdotal evidence and few actual comparative studies of the two school structures (Balfanz 2002; Coladarci and Hancock 2002; Hough 2005; McEwin et al. 2005; Pardini 2002; Weiss and Kipnes 2006; Yakimowski and Connolly 2001). That is where this study hopes to make a significant contribution to the field. By employing a more appropriate method of statistical analysis, a substantially larger sample size, and a more diverse set of statistics controls, we provide a much stronger scientific analysis comparing the mathematics and reading achievement levels across middle and K–8 schools.

Of the research on K–8 versus middle schools directly cited in this study, only the studies by Offenberg (2001), Simmons and Blyth (1987), and Weiss and Kipnes (2006) employed rigorous statistical analyses. Weiss and Kipnes (2006) used a comparative sample and multilevel modeling and found that students at K–8 schools reported higher levels of self-esteem and felt less threatened at school. Offenberg (2001) employed a school-level analysis, finding some achievement advantages for students in K–8 schools, such as higher standardized test scores and better grade-point averages. The last, by Simmons and Blyth (1987), was a student-level analysis that found students from K–8 schools to enjoy higher levels of social engagement, better attitudes toward school, and also higher levels of self-esteem.

In this study we take a step forward and merge the above techniques by using multilevel modeling for our analyses (Bryk and Raudenbush 1992; Snijders and Bosker 1999). Multilevel models, or hierarchical linear models, ac-

count for grouped data such as ours, where students are nested within schools. Multilevel modeling is similar to regression modeling but takes into account the fact that, with nested data, students within the same school will have shared similar experiences and thus they will not be independent of each other, violating a statistical assumption of standard regression modeling. This difficulty was the very reason why Offenberg (2001) decided to focus on the school level for analysis, but here we will be able to capture both the student- and school-level factors that influence student achievement in a statistically appropriate fashion.

Our models are in fact three-level models and also include an intermediate level between students and schools to account for the cohort in which the students are nested, with the cohorts themselves nested within schools. This allows us to take into account the fact that our data are spread over several years and that within each school, different cohorts of students may vary from year to year in significant ways that affect student achievement.

Sample

Our sample runs from the 1999–2000 school year to the 2003–4 year and covers 40,883 eighth-grade students, taken from 95 schools over the five years. While the number of students represents the sample size for level 1 of our model (between students) and the number of schools is our level 3 sample size (between schools), our sample at level 2 (between cohorts) is 427, which represents the number of cohorts we have across the five-year time span for our 95 schools. The number is not exactly equal to 475 (95 schools × 5 years) because of the changing grade structures of the schools in our sample as Philadelphia implemented its K–8 conversion policy. For example, during the five years observed in our study, 14 elementary schools were transformed into new K–8 schools and so did not have eighth grades throughout all the years of analysis. These elementary schools added one grade per year, and of these 14 new K–8 schools, one reached eighth grade in spring 2000, four in spring 2001, two in 2003, and seven added the eighth grade in spring 2004. Conversely, six middle schools were being transformed into higher-level schools containing grades 5–12 or 6–12, junior highs, or high schools. Their transformations began in the 2003–4 year, at which point their cohorts were excluded, as the schools no longer fit into the middle school versus K–8 comparison. Missing data also cost us the inclusion of one cohort in one school, and finally one middle school ceased to exist entirely as part of the school district after the 2000–2001 school year.

Of our K–8 schools, one was actually a grades 1–8 school, and of our middle schools, 17 were grades 5–8, 20 were grades 6–8, and two were

originally grades 6–8 but were transformed into grades 7–8 during the period under observation. No K–12 schools were examined, as they fell outside the domain of the policy of converting K–8 into middle schools and thus were of no aid in testing our direct hypotheses comparing the benefits of those two school types. Five schools were also left out of our analysis entirely due to their unique and substantively different nature in comparison to the typical middle-grades schools. Three of these were schools that accepted their student body on a selective basis, one was a year-round school, and the last was not a local neighborhood school but rather took in its students on an application and lottery basis. In the end, we were left with a total of 39 middle schools to which we could compare 42 old K–8 schools and 14 newly formed K–8 schools.

Measures

Another advantage of this study is that it captures and combines the majority of statistical controls that are considered to be of theoretical relevance, whereas previous studies have been able to focus only on a select few. This allows us to compare their relative impacts on any K–8 and middle school differences, while at the same time gain a better understanding of any such differences as a whole. All our data were received directly from the Philadelphia School District and given as secondary copies of existent data previously collected by the district for their own use.

The Outcome

In all our analyses, our outcome measure was students' eighth-grade scores on the Pennsylvania State System of Assessment (PSSA). The particular metric used is normal curve equivalents (NCE), which are similar to percentiles but equidistant along a normal distribution curve, making the difference between the first and second NCE equivalent to the distance between the forty-ninth and fiftieth, unlike with percentiles. Normal curve equivalents are superior to scale scores, percentiles, and grade equivalents for the purposes of summary statistics, and gain scores and were originally designed specifically for use in education research and evaluation.

Using the PSSA as an outcome is highly appropriate as it is the "high stakes" test used by the state to evaluate schools and districts and assess their annual performances. The test results are one of the key measures of accountability that schools are held to, especially since the introduction of the No Child Left Behind (NCLB) federal legislation act. As NCLB has substantially increased the

emphasis placed on test scores and the percent of students scoring below basic, the importance of test performance has grown dramatically, as has its impact upon the day-to-day activities of teachers and students. The difference between good and bad yearly results can mean the difference between levels of funding and affect the futures of school staff and administrators.

Prior Achievement

Also included were the students' fifth-grade scores on the PSSA, used to control for their prior levels of achievement (fifth grade is the last year in which the test was administered prior to the eighth grade). If K–8 schools are in fact better for student learning and achievement, students at K–8 schools may already have higher levels of prior achievement by the end of the fifth grade, as several middle schools begin in the fifth grade and their students have already made the transition to new middle schools. It is thus important to control for students' prior achievement, especially in this quasi-experimental setting, and such a covariate substantially increases the power of our models (Bryk and Raudenbush 1992; Snijders and Bosker 1999). Also, as our focus is upon the differing abilities of the two school structures to contribute to students' academic development during the middle grades alone, controlling for any prior differences allows us to focus on this.

Time

We examined the use of different measures to account for both change over time in student test scores and changes between cohorts. We found that each cohort had improved its average PSSA scores over those of its predecessors. A continuous measure for time accounted for this growth with statistical significance, and, controlling for time, only the final cohort was significantly different from the others, as it scored significantly higher than the others on average, even after controlling for the yearly trend upward.

This pattern over time and across cohorts is likely to be partially due to schools, teachers, and students becoming more familiar with the test over time, along with increased teaching to the test and greater alignment of school curriculum to match the contents of the test. The PSSA was piloted in spring 1996 and given district-wide for the first time in the 1996–97 school year. It also grew in importance in later years, as it was not used as the main method of state and district evaluation for schools until roughly the turn of the century, when another test that had previously served as the main measure of school evaluation, the SAT-9, was phased out.

Student Demographics

In terms of student population demographics, at the student level we have a dichotomous variable for gender, dummy variables for the ethnicities of Asian, Hispanic, black, white, and "other," as well as dichotomous measures for special education status and English as a second language (ESL) status. At the cohort level we have a measure for the percent of students in each cohort that were eligible for the free/reduced lunch program (FRL). While FRL status is truly a student-level factor, the data are not released for individual students and thus can be included only at the cohort level. Another measure captures the percent of students in each cohort that were Hispanic or black. Asians are not included when aggregating the proportion of minority students in each cohort as they typically perform much better in terms of academic achievement than other ethnicities (Kao 1995; Peng and Wright 1994), a decision with precedents in the literature (Offenberg 2001).

Teacher Data

For teacher characteristics we included several measures for various teacher qualities, all aggregated to the cohort level. One was a measure for teacher absentee rates (average percent of contractual days missed by all teachers at the school), another was the percent of certified teachers at the school (certified by the Pennsylvania Department of Education), a third was the average experience of the teachers at the school (measured by the average number of years that teachers had been registered in the Philadelphia School District), and a fourth was the student/teacher ratio of each school (not the most reliable measure of class size and typically an underestimate but the only one available to us).

One more measure for teacher characteristics was left out of our final models due to missing data. The percent of teachers returning from the previous year (retention) was available for all cohorts except the first, 1999–2000. However, exploratory analyses were run with the subsample of our data for which teacher retention was available, and no significant correlations between the percent of returning teachers and students' achievement scores were found, nor did teacher retention have an impact upon any differences between K–8 schools and middle schools. Descriptively, we also found that only new K–8 schools had significantly lower rates of returning teachers, and this was due to their K–8 conversion and the expanding grade levels that led to a nominal increase in the number of teachers and a proportional increase in the percent of new teachers at the schools.

School Transition

Furthermore, and of great theoretical importance to us, by using the district's administrative records we were able to create a dichotomous variable to control for whether or not students were in the same school in eighth grade as they had been in the fourth grade of elementary school. This provided us with a proxy for school transition through which we could determine if making a transition to a new school in the middle grades, as all middle school students must do, was detrimental to student achievement and a factor in any differences between K–8 and middle schools.

Though this measure is an important difference between middle schools and K–8 schools, it is a characteristic of each individual student and so is a level 1 variable in our models. It also does not correlate perfectly with attending a K–8 versus a middle school since not all K–8 students attended the same school in grade 8 that they did in grade 4, and many K–8 students changed schools for reasons other than attending a middle school.

School Factors

At the cohort level we included three other measures that were separate from student demographics and teacher characteristics. While only school size is of direct theoretical relevance as per the literature, all were of some interest and value in estimating student achievement. Included is a dichotomous variable that takes into account whether a school was under a new principal for each cohort's eighth-grade year, and another measure that accounts for school-level mobility rates during the school year (the proportion of a school's enrolled students who either came after the start of the school year or left before the end). This last measure (mobility) is a level 2 variable and a cohort-level characteristic measured during each school year, not to be confused with our level 1 variable (transition) for students who transitioned between elementary and middle grades, as only the latter is of theoretical relevance to our analysis.

School size was first measured in three different ways. In exploratory analyses, we tried measures for the size of the eighth grade alone, the average size of the middle grades (5–8), and the average size of all grades in the school. The third and final measure, average grade size, proved to be the most significantly correlated to student achievement, both with and without other controls in place, and thus was kept as the measure for our final analyses. We believe this to be because overall grade size within the entire school, more than just a student's own grade, or just the middle grades of a school, is what has the most effect upon the quality of a student's learning environment.

At the third level of our model, the between-school level, we included only

two distinct measures. The first was geographical region, split into eight dummy variables for the nine local regions recognized by the school district. Prior research has suggested region as another possible cause of aggregate differences between K–8 schools and middle schools, since a larger proportion of the K–8 schools in Philadelphia are found in areas of higher socioeconomic status (Balfanz 2002). Also, due to the rollout of the policy in Philadelphia, which saw the creation of new K–8 schools by region, newly converted K–8s are disproportionately represented in certain geographic regions.

The second school-level factor, school structure, is our most important variable of all, at least so far as our theoretical interests are concerned. This was measured by three mutually exclusive and exhaustive dummy variables that compared both old K–8 schools and newly formed K–8 schools (created in the last five years as part of the district's reform policy) to middle schools. It is through them that we are able to determine if any differences in student achievement exist between K–8 schools and middle schools and to what degrees any such differences might be affected by controlling for population demographics or school structure. By comparing the older K–8 schools and the newer ones separately, we are also able to see if the older schools did indeed have a student achievement advantage compared to middle schools and then if the district has been able to replicate any such advantage in the newer K–8 schools with its conversion policy.

Descriptive Comparison

Tables 1–2 show the results from *t*-tests comparing the different school structures along the above-mentioned variables. Both old K–8 and new K–8 schools are compared separately to middle schools, allowing us to see the descriptive differences between existing K–8 and middle schools and then also the differences between middle schools and the newer K–8 schools that have been created as part of the district's reforms.

Between the old K–8 schools and middle schools, we see significant differences along all the theoretically presumed dimensions. The old K–8 schools show significantly higher levels of achievement in both the fifth and the eighth grades. At the same time, they show significantly lower proportions of Hispanic, black, and high-poverty students compared to middle schools, but significantly higher proportions of white and Asian students. The majority of students at K–8 schools were in the same school at grade 8 as in grade 4, and in general they had much smaller average grade sizes and experienced much lower rates of student mobility during the year. In addition, K–8 schools had teaching staffs that averaged more than three years greater experience compared to teachers at middle schools, while also having lower rates of

TABLE 1

Student-Level Descriptives

Variable	Middle School (N = 28,595)	Old K–8 (N = 10,938)	New K–8 (N = 1,350)
Grade 8 math score	34.0	41.9***	33.1*
	(16.6)	(17.0)	(14.7)
Grade 5 math score	29.9	35.4***	24.7***
	(17.3)	(17.4)	(14.9)
Grade 8 reading score	34.5	42.1***	33.9
	(16.6)	(17.4)	(14.3)
Grade 5 reading score	30.2	35.4***	26.4***
	(16.8)	(17.8)	(15.0)
Female (%)	53	53	53
Special education (%)	15	24***	13*
ESL (%)	5	5	6
White (%)	13	27***	1***
Black (%)	71	53***	72
Asian (%)	4	9***	2***
Hispanic (%)	11	10***	25***
Other ethnicity (%)	< 1	< 1	< 1
Same school in grades 4 and 8 (%)	0	63***	63***

NOTE.—Numbers in parentheses are standard deviations.
* Significant at .05 level.
** Significant at .01 level.
*** Significant at .001 level.

teacher absenteeism and greater proportions of certified teachers. All the above differences between old K–8 schools and middle schools were significant at the $\alpha - .001$ level, confirming that K–8 schools did indeed have higher levels of student achievement but at the same time served significantly differently student populations with much lower proportions of minority students from high-poverty backgrounds, who experienced a low rate of transition to new schools, on average attended much smaller schools, and were taught by teachers with greater experience, better attendance, and higher rates of certification.

The differences between the new K–8 schools and middle schools are quite interesting in terms of our theoretical variables and the implications for evaluating the district's K–8 conversion policy. The students at the 14 new K–8 schools created during the time period of our analyses, like students at the old K–8 schools, experienced fewer transitions to new schools and, on average, were in much smaller schools when compared to middle school students. However, the 14 schools selected for conversion by the district actually served student populations with lower levels of achievement than middle school stu-

TABLE 2

Cohort-Level Descriptives

Variable	Middle School ($N = 187$)	Old K–8 ($N = 208$)	New K–8 ($N = 32$)
New principal (%)	25	21	34
Free/reduced lunch program (%)	79.7	68.0***	93.3***
	(16.6)	(21.5)	(5.1)
Minority (Hispanic + black) (%)	85.6	64.8***	96.3***
	(20.1)	(24.8)	(9.7)
Average grade size	248	74***	76***
	(85.0)	(30.4)	(22.5)
Student mobility	37.5	30.9***	44.2**
	(11.5)	(11.9)	(7.8)
Mean grade 5 math score	29.2	34.9***	24.9**
	(7.6)	(7.6)	(6.1)
Mean grade 5 reading score	29.7	34.8***	26.5***
	(6.8)	(7.3)	(4.2)
Teacher absentee rate (%)	6.6	5.7***	6.5
	(2.0)	(2.1)	(1.9)
Certified teachers (%)	85	93***	81*
Teacher experience	11.7	14.8***	9.5***
	(3.3)	(4.4)	(2.6)
Student/teacher ratio	18.1	17.6*	17.3
	(2.5)	(2.6)	(2.2)
Cohort 2000 (%)	21	20	3***
Cohort 2001 (%)	21	20	16
Cohort 2002 (%)	21	20	16
Cohort 2003 (%)	20	20	22
Cohort 2004 (%)	17	20	44**

NOTE.—Numbers in parentheses are standard deviations.
* Significant at .05 level.
** Significant at .01 level.
*** Significant at .001 level.

dents and with larger proportions of Hispanic students and fewer Asian and white students. The teachers at the new K–8 schools were also significantly less experienced than their middle school counterparts. Thus, while the new K–8 schools had the advantages of smaller grade sizes and low school transition, they actually served student populations with greater percentages of minority students who on average had lower levels of achievement when compared to middle schools and who were taught by less experienced teachers.

Looking at changes over time, we also see, as expected, the effect of the district's ongoing reforms in terms of the sample sizes for each school type. The number of students at middle schools decreases by the 2003–4 and final

cohort, while the number of students attending K–8 schools increases. In 1999–2000, 73 percent of that cohort's students were enrolled in middle schools and 27 percent in K–8 schools. By 2003–4, with the introduction of the 14 new K–8 schools and the phasing out of some middle schools, those same numbers were reduced to 64 percent of students enrolled in middle school versus 36 percent at K–8 schools as we see the district's K–8 conversion policy taking effect.

Hypotheses

Moving forward to our multilevel model analysis, the question of comparing middle schools to K–8 schools has become more intricate. Both the old K–8 and the new K–8 schools share the intrinsic advantages over middle schools of smaller sizes and low school transition rates. However, we see that while the old K–8 schools have more advantaged student and teacher populations, the new K–8 schools in fact have more disadvantaged populations than do the middle schools.

This leads us to put forth the following hypotheses: (1) as the old K–8 schools are more advantaged in terms of both the external and intrinsic qualities, they should have a significant advantage over middle schools in terms of student achievement; (2) since new K–8 schools have intrinsic advantages over middle schools but at the same time serve more disadvantaged populations, they should not perform significantly differently from middle schools in the end; (3) if we control for the external factors, student and teacher characteristics, any old K–8 advantage over middle schools should be reduced, while new K–8 schools should improve in comparison to middle schools; (4) if we control for both the external and intrinsic qualities, there should be no significant differences between either old or new K–8 schools and middle schools.

Model Building

While there are several different techniques for building regression models and even more for constructing multilevel models, we chose to build ours in a series of four steps, guided by theoretical interest, as follows. This process was done first for mathematics achievement as an outcome and then again for reading. Broadly, the four steps began first with an empty model and then, second, added in measures for time, cohort, and prior achievement, followed by population demographics, third, and then, last, the measures intrinsic to school structure and K–8s.

Before and after each step, we added in and then removed the dummy variables for old and new K–8 schools. This model-building process allowed us to first see what, if any, differences existed between these school types and middle schools in an empty and unconditional model; second, how student achievement growth was shaped by time, cohorts, and prior achievement; then, third, how the differences between K–8 and middle schools were affected by the inclusion of demographics; and, fourth, how any such differences were affected by the inclusion of measures for school structure. Finally, after the inclusion of all our measures, we were able to see if there remained any significant differences between school types at all.

This technique also overlaps largely with the idea of building our models from the first level up (Bryk and Raudenbush 1992), starting with the between-student-level variables and then moving on to cohort- and then school-level variables. Along the way and separately at each step, we removed any variables that were not statistically significant one at a time, moving in a backward fashion and starting with the least significant estimates. This kept with the idea of valuing parsimony among other things, as multilevel models can become quite complex, especially in the reporting of large and complicated models. We wished to keep our focus to our theoretical concerns and to keep our statistical reports succinct and of the most value and interest to readers.

After coming to our final results, we then also rebuilt our models using other strategies as a check. Each time, we arrived at similar models with comparable results that found impacts of similar magnitudes for our variables of interest. The models were all run using HLM 6.0 software and the REML method of estimation.

Analysis

Our results begin with our empty models and, prior to adding any control variables, studying the variation in our outcome. Given our sample, we find that the random variation in our outcome is highly significant at the $\alpha - .001$ level for all three levels of our model, given the chi-square statistics provided by HLM software (Bryk and Raudenbush 1992), in both our math and reading models. Thus there is separate and unique variation between students, cohorts, and schools in terms of achievement and empirical justification for the use of our three-level models. We also find that 76 percent of the variation in math achievement is between students at level 1 of our model, 6 percent between cohorts at level 2, and the remaining 18 percent varying between schools at level 3. For reading, the numbers were 79 percent, 4 percent, and 15 percent. This is consistent with other research on school effects that has found between

10–30 percent of the variability in student achievement is typically between schools (Bryk and Raudenbush 1992).

Our first step after running an empty model was to add in dummy variables for new and old K–8 schools, in order to determine if these school structures did indeed have any significant differences in terms of achievement, prior to controlling for any possible causes. As prior studies and descriptive analyses have found, old K–8 schools did indeed have a large and significant advantage over middle schools of over 8 NCE on both the mathematics and reading parts of the PSSA exam (math: $\beta = 8.55$, $t^* = 5.89$, $p < .000$; reading: $\beta = 8.23$, $t^* = 5.70$, $p < .000$). New K–8 schools, however, did not have statistically different averages from middle schools in either subject.

Before trying to explain the differences between school structures, we first added in controls for time. As mentioned above, we found that each cohort had improved over its predecessors in linear fashion. A continuous measure for time accounted for this trend and saw each cohort grow by approximately 1 NCE more than the previous cohort in both math and reading. Even after accounting for the linear increases over time, the final cohort of 2003–4 still outperformed all other cohorts by just over 1 NCE in our final models for both subjects. However, while statistically significant and necessary in terms of model specification, neither of our measures for time and cohort affected the differences between school structures.

Prior achievement, or fifth-grade score, was then added to our model at the student level and also at the cohort level as an aggregated measure, though only the level 1 measure proved to be significant and was maintained throughout later modeling. Prior achievement at the first level was also grand-mean centered, making its estimate one for a student with the average level of prior achievement. The effect of adding prior achievement into the model was to halve the difference between old K–8 schools and middle schools (math: $\beta = 3.60$, $t^* = 5.12$, $p < .000$; reading: $\beta = 3.97$, $t^* = 6.24$, $p < .000$). However, this merits some qualification as it does not so much halve the effect of K–8 schools as halve the period of time over which we are comparing them. By adding prior achievement into the model, we are finding that most of the difference between K–8 students and middle school students is established by the end of grade 5. Controlling for their achievement at grade 5 and the distance by which K–8 students are already ahead at that point, we see that K–8 students were still scoring over 3 NCE higher on average on the eighth-grade test across subjects. The t^* statistic and p-values for the difference between K–8 schools and middle schools remain largely unchanged by the addition of prior achievement as a factor.

Moving past our basic statistical controls, we come to our set of variables for student demographics. At level 1, the between-student level, we found that all of our demographic variables were significant except for "other" ethnicity,

which consisted of less than 1 percent of our student sample. Female, Hispanic, white, and Asian students all scored significantly higher then black and male students on average. Special-education students also scored approximately 0.5 and 1 NCE lower than non-special-education-status students in math and reading, respectively, while ESL students scored roughly 1 and 2 NCE lower across math and reading. At the cohort level, the percent of students eligible for the free/reduced lunch program was statistically significant but the percent of minority students in a cohort was not. The effect of FRL was to reduce a cohort's average for achievement by approximately 0.4 NCE in math and 0.3 NCE in reading for each additional 10 percent of students eligible for the FRL program, controlling for our other factors.

The effect on the K–8 advantage of controlling for all our statistically significant student demographic factors was to reduce its estimate by approximately 0.5 NCE in math and 1.5 NCE in reading (math: $\beta = 3.23$, $t^* = 4.31$, $p < .000$; reading: $\beta = 2.34$, $t^* = 4.11$, $p < .000$), while keeping the new K–8 schools statistically similar to middle schools.

Our measure for teacher absentee rates was not significant in either our math or reading models, but teacher-student ratios were significant in both sets of models, and teacher experience was significant in regards to reading achievement alone. While the old K–8 advantage in math was not affected by the additions of teacher quality controls ($\beta = 3.18$, $t^* = 4.13$, $p < .000$), there was a substantial impact on the reading achievement advantage ($\beta = 1.75$, $t^* = 3.23$, $p < .002$). Controlling for teacher characteristics also brought the difference between new K–8 and middle schools to a level of significance in reading ($\beta = 1.61$, $t^* = 2.07$, $p < .041$), but not in math.

Next we moved from the external factors to the intrinsic ones and added our student-level measure for school transition, or rather for staying in the same school from elementary to middle grades. This estimate was statistically significant and reduced the effect of old K–8 schools on average math achievement by 1 NCE ($\beta = 1.68$, $t^* = 2.17$, $p < .033$), while pushing its statistical significance above the $\alpha - .010$ level. In reading, the old K–8 school advantage was also reduced by 1 NCE ($\beta = 0.65$, $t^* = 1.09$, $p < .278$) and no longer significant. The effect of including school transition on the new K–8 reading advantage was the same, reducing it by roughly 1 NCE ($\beta = 0.52$, $t^* = 0.63$, $p < .531$) and making it nonsignificant. Overall, students who were in the same school in grade 4 as in grade 8 scored on average almost 2 NCE higher in both math and reading than students who transitioned in the middle grades.

Of our remaining cohort-level factors, our measures for average grade size and student mobility proved to be statistically significant while our variable for new principals did not. Average grade size had the single largest effect on the K–8 advantage, reducing the estimated coefficient for the effect of old

K–8 schools on average mathematics achievement to statistically no different from zero ($\beta = 0.15$, $t^* = 0.15$, $p < .881$) and further reducing the old K–8 advantage in reading ($\beta = 0.54$, $t^* = 0.57$, $p < .567$). The differences between new K–8 schools and middle schools were also reduced both nominally and in terms of statistical significance in both math and reading. The overall effect of grade size was to decrease a cohort's average achievement score by roughly 0.8 to 0.4 NCE per each additional 100 students per grade, in math and reading, respectively. As the average difference in grade size between our K–8 and middle schools was a difference of 173 students, this translates into an effect of 1.4 and 0.7 NCE in the difference between the average cohort scores from K–8 and middle schools, controlling for other factors, for math and reading, respectively. (We also found no evidence of an interaction between grade size and K–8 schools as was found in Offenberg's [2001] study). Student mobility, while in and of itself significant in its impact upon student achievement, did not have a substantial effect on the differences between school structures.

Finally, we included our third-level dummy variables for region, though they were not statistically significant predictors of student achievement. Alone, without other control variables, there were some significant differences between regions. However, after controlling for student demographic factors, these few differences disappeared, as it is largely population differences that seem to drive any regional disparities. Including region also did not affect either of our variables for school structure.

Table 3 highlights the estimated coefficients for our old K–8 and new K–8 measures at each stage of our model-building process, after the inclusion of various controls, and table 4 shows the estimates for the rest of our measures from our final models for both subjects. Compared to our earlier empty model, approximately 46 percent of the total variation in mathematics achievement was explained by the explanatory variables in our final model. Forty-five percent of the variation between students was explained, along with 65 percent of the cohort-level variation and 44 percent of the variation between schools. In reading, our model led to a proportional reduction of 46 percent in the total variation for reading achievement, 44 percent of the between-student variation, 63 percent of the cohort variation, and 53 percent of the variation between schools. Comparing the deviance statistics (-2 log likelihood) of our empty and final models, we arrive at chi-square test statistics of 23,947.2 with 55 degrees of freedom (*df*) for math and 23,064.9 with 62 *df* for reading. Both result in highly significant *p*-values at the $\alpha - .001$ level, confirming that the explanatory variables included in our final models have in fact contributed statistically significant information regarding our outcome variables, students' eighth-grade achievement scores.

Strengths and Limitations

Statistical Conclusions

Before moving on, it is best to first highlight some of the strengths and limitations of this study, as they shape the validity and reliability of any conclusions we might draw. That student achievement varies randomly at both the cohort and school level represents empirical evidence that multilevel models are in fact an appropriate and necessary method for analyzing the research questions presented in this article. This is further supported by the observance that the slopes of several of our level 1 student predictors also varied randomly at the cohort and school levels.

Completeness of Data

In this study, the large samples upon which our models were based, taken over a long period of time, provide great strength to our estimated parameters. As this is a natural experiment comparing K–8 schools to middle schools in the Philadelphia City School District alone, our sample of schools represents the entire population of local neighborhood public schools in the district. At the cohort level, we also had data for all but one of the cohorts to pass through the schools during the period under observation. However, while our study included virtually all schools and cohorts, it included only a proportion of the eighth-grade students to have passed through the district during this time.

As our set of statistical controls was rather robust and one of our main strengths, it also meant that we required a great deal of data for each student that we included. We relied only on cases for which we had complete data, and thus in the end many students had missing data and were left out of the analyses. For example, we might have had one student's eighth-grade score and all their demographic data but lacked their fifth-grade score for prior achievement or fourth-grade record for prior school.

To assess any bias that might exist in our sample, we were able to estimate the percentage of students for which we had complete data and therefore the percentage of each cohort that was included in our analyses. This ranged from a high of 95 percent to a low of 24 percent. However, only 8 percent (34 of the 428 cohorts) were below 50 percent representation, and, conversely, 92 percent of cohorts were represented by half of their students or more. Sixty-eight percent of the cohorts were over 60 percent representation, 28 percent over 70 percent, 7 percent over 80 percent, and 3 percent over 90. The representation of students was slightly higher in K–8 schools where at-

TABLE 3

Coefficient Estimates for Old K–8 and New K–8 Schools (Compared to Middle Schools)

	MATHEMATICS				READING			
	β	t^*	p-value	95% CI	β	t^*	p-value	95% CI
Old K–8:								
With no statistical controls	8.56			11.4	8.23			10.9
	(1.45)	5.89	.000***	5.7	(1.37)	6.01	.000***	5.6
After including measures for time & cohort	8.41			11.2	7.97			10.7
	(1.42)	5.93	.000***	5.6	(1.41)	5.64	.000***	5.2
After controlling for prior achievement	3.60			5.0	3.97			5.2
	(.70)	5.12	.000***	2.2	(.64)	6.24	.000***	2.7
After adding student demographics	3.23			4.7	2.34			3.5
	(.75)	4.31	.000***	1.8	(.57)	4.11	.000***	1.2
After adding teacher characteristics	3.18			4.7	1.75			2.8
	(.77)	4.13	.000***	1.7	(.54)	3.23	.002**	.7
After including school transition	1.68			3.2	.65			1.8
	(.77)	2.17	.033*	.2	(.59)	1.09	.278	−.5

After controlling for average grade size	.15			2.1	.54			2.4
	(.97)	.15	.881	−1.8	(.95)	.57	.567	−1.3
New K–8:								
With no statistical controls	1.16			4.8	.67			4.7
	(1.84)	.63	.529	−2.5	(2.06)	.326	.745	−3.4
After including measures for time & cohort	−1.40			2.0	−1.41			1.3
	(1.74)	−.80	.425	−4.8	(1.39)	−1.02	.312	−3.9
After controlling for prior achievement	1.06			3.3	.94			2.7
	(1.16)	.92	.362	−1.2	(.90)	1.04	.301	−.8
After adding student demographics	1.68			3.9	1.49			3.1
	(1.15)	1.46	.144	−.6	(.82)	1.81	.073	−.1
After adding teacher characteristics	1.79			4.0	1.61			3.1
	(1.15)	1.56	.122	−.5	(.78)	2.07	.041*	.1
After including school transition	.90			2.7	.52			2.1
	(.92)	.98	.330	−.9	(.82)	.63	.531	−1.1
After controlling for average grade size	−.24			1.8	.49			2.5
	(1.05)	−.23	.819	−2.3	(1.03)	.40	.632	−1.5

NOTE.—CI = Confidence interval.
* Significant at .05 level.
** Significant at .01 level.
*** Significant at .001 level.

TABLE 4

Final Model Fixed Parameter Estimates

	MATHEMATICS			READING		
Fixed Effect	Coefficient	SE	*p*-value	Coefficient	SE	*p*-value
Intercept, G000	38.97	2.22	.000***	36.31	2.37	.000***
% FRL, G010	−.04	.02	.019*	−.03	.01	.029*
Mobility, G020	−5.56	2.31	.018*	−5.58	2.20	.012*
S/T ratio, G030	−.13	.08	.119	−.19	.08	.022*
Experience, G040	.04	.09	.636	.22	.07	.003**
Grade size, G050	−.008	.003	.004**	−.003	.002	.059
Time, G060	.99	.17	.000***	.62	.14	.000***
Cohort 99–00, G070	1.64	.45	.001***	1.29	.37	.001***
Slope for prior achievement, fifth-grade NCE:						
Intercept, G100	.60	.01	.000***	.60	.01	.000***
Slope for gender:						
Intercept, G200	.59	.14	.000***	2.31	.13	.000***
Slope for special education:						
Intercept, G300	−.42	.28	.132	−1.02	.32	.002**
Slope for ESL:						
Intercept, G400	−.98	.47	.038*	−2.13	.40	.000***
Slope for white:						
Intercept, G500	1.11	.24	.000***	1.09	.27	.000***
Slope for Asian:						
Intercept, G600	7.17	.34	.000***	5.63	.37	.000***
Slope for Hispanic:						
Intercept, G700	.73	.19	.000***	.73	.20	.000***
Slope for same school in grades 5 and 8:						
Intercept, G800	1.61	.32	.000***	1.94	.25	.000***

NOTE.—SE refers to standard error, FRL refers to free/reduced lunch program, ESL is English as a second language, and NCE is "normal curve equivalents."

* Significant at .05 level.

** Significant at .01 level.

*** Significant at .001 level.

tendance is higher, where on average 67 percent of the students from each cohort were included compared to only 60 percent for middle schools. Furthermore, we were able to compare the true cohort averages for our outcomes, eighth-grade PSSA scores, to our sample averages, though only in terms of scale scores and not the NCE metric used in the study. Since we knew that using only complete cases is a biased and inefficient way of dealing with missing data, this then allowed us to determine the nature of any bias inherent in our

analysis. We found that while the true mean score for the cohorts was a scale score of 1,183 in math and 1,168 in reading, our sample means were 1,197 in math and 1,192 in reading, thus overestimating by differences of 14 and 24 scale scores, respectively.

Combining the above facts, we would hypothesize that those students who are missing from our analyses are likely those who are the most transient and those with the lowest achievement levels who may have repeated grades. Transient students would be the most likely to be out of school on test day, have missing or unrecorded data, and likely have lower levels of academic achievement than the rest of their cohorts for whom we have data. We also found that while the true mean attendance rate for our cohorts was 89.4 percent of school days, our sample mean was also overestimated at 91.6, thus confirming this hypothesis. Repeating students are also the most likely to be lost from our sample, as repeating a grade makes it more difficult to track an eighth-grade student back to their fifth-grade prior score and their fourth-grade prior school as they fall out of their original cohort. In addition, the nature of multilevel models, which require each lower-level unit to be nested within only one higher-level unit, also presents a problem as students who repeat a grade would then have membership in two different cohorts. Still, considering the high degree to which the cohorts are represented in our sample, and the small magnitude of the difference between our sample means and the true cohort means for our outcomes, we believe that our large sample and also our model estimates are overall highly representative of the true population.

As opposed to missing cases, the issue of missing variables also sets some limitations on our study. While with our abundant number of measures we were able to analyze the academic differences between K–8 and middle schools in great depth, a particular lack of social engagement and attitudinal measures in our data limited the detail of our explanation for those differences. For example, while we know that grade size and school transition are significant contributors to the differences in achievement performances between the two school structures, our models were not able to show how they do so through the promotion of students' relationships with each other and staff, by increasing their self-esteem, and providing greater involvement in leadership and extra-curricular activities. These types of measures typically come from surveys and self-reported means that were not available to us here. However, these social engagement aspects have been addressed in much of the past research (Simmons and Blyth 1987; Weiss and Kipnes 2006), and this article serves them well as a complement by estimating the achievement aspect and isolating its causes in a sound empirical manner.

Measurement Issues

In terms of construct validity, and while our robust set of statistical controls are one of this study's strengths, there are still some comments that need to be made in order to provide the appropriate context for our results. One comment relates to our measures of teacher characteristics, and we would not presume that, in regard to these factors at least, our study provides any conclusive results. As all our teacher measures were aggregated to the cohort level, we would think that our results might suffer from some problems of construct validity. With achievement outcomes for individual students in their eighth-grade year, we would hypothesize that what is most relevant to these students, and to our models, would be these same concepts measured at the individual level for their own actual eighth-grade math teachers. With disaggregated teacher measures, such factors might have achieved higher levels of significance in our models.

Also, our measure for prior achievement is taken in grade 5, after many students have already entered into middle schools. This then reduces the validity of our results, though only to a small extent, as it is a small proportion of students of which we speak and our models still isolate grades 6 through 8 and the majority of their middle-grade years.

Additionally, our measure of school transition cannot measure the "Top Dog" theory introduced by Simmons and Blyth (1987). Since no middle school students were in the same school in grade 4 and grade 8 (by definition), and since no K–8 students switched to a middle school for the middle grades (or they would be classified as middle school in our model), we then do not have available to us in the data the counterfactual needed to evaluate this hypothesis. However, this theory is not the only hypothesis of why changing schools may hurt student achievement, nor is it the primary one. The same authors point to the adjustments to new staff and students and the new relationships that a student must make when changing schools as having a major effect on student achievement. This occurs even for K–8 students who changed to other K–8 schools, an effect that our data allows us to measure and estimate.

Finally, while our measures for grade size and school transition are highly correlated to our dummy variables for school structure, we are not concerned with collinearity for two reasons. First, the variables are not perfectly correlated, as for school transition not all K–8 students remained in the same school from grade 4 to grade 8, and for grade size there is overlap between the range in sizes for the two school structures, with some middle schools under 100 students per grade and some K–8s above that. Second, while school structure is a level 3 variable and alternates with each observed school, grade size is a level 2 variable observed with each change in cohort, and school transition a level 1 variable measured for each individual student. Furthermore, it is

one of the essential points of this article that school structure is highly correlated to both of these measures, and that differences between the school types in achievement are highly attributable to their differences in grade size and school transition. Besides estimating any achievement differences between the two school types, we have also sought to adequately explain why such differences exist, and while some reasons, such as student and teacher demographics, are external to the school structures, two other reasons are the smaller size and greater continuity that are largely intrinsic to K–8s and synonymous with the reform.

External and Internal Validity

Despite these qualifications, we still believe that this study has a high level of internal validity, a high power to detect any true effects of school structure upon student achievement, and that it provides relatively unbiased estimates of the true population parameters that we have examined. This strength, again, is based upon the method of analysis used; the large representative samples of schools, cohorts, and students incorporated over a wide time frame; and the large set of theoretically relevant statistical controls that we have been able to include.

Beyond this study, and in regard to external validity, we believe that the Philadelphia City School District makes an excellent case study from which to generalize to other large urban districts. It represents one of the largest urban public school districts in the United States and serves a student population that consists largely of minority students from high-poverty backgrounds. Combined with the fact that it has already enacted a policy of K–8 conversion, it makes an ideal case in which to study the effects of such a conversion, and for other districts that are considering such a policy to learn some early lessons from.

Discussion

In terms of our earlier hypotheses, we have seen the following: (1) old K–8 schools with both external and intrinsic advantages did, in fact, have significantly higher levels of achievement; (2) between their more disadvantaged student and teacher populations and their intrinsic advantages over middle schools, newer K–8 schools did not perform statistically differently in terms of student math and reading achievement; (3) after controlling for the external factors of population demographics, the old K–8 advantage was reduced, though still significant, while the new K–8 schools developed a statistically

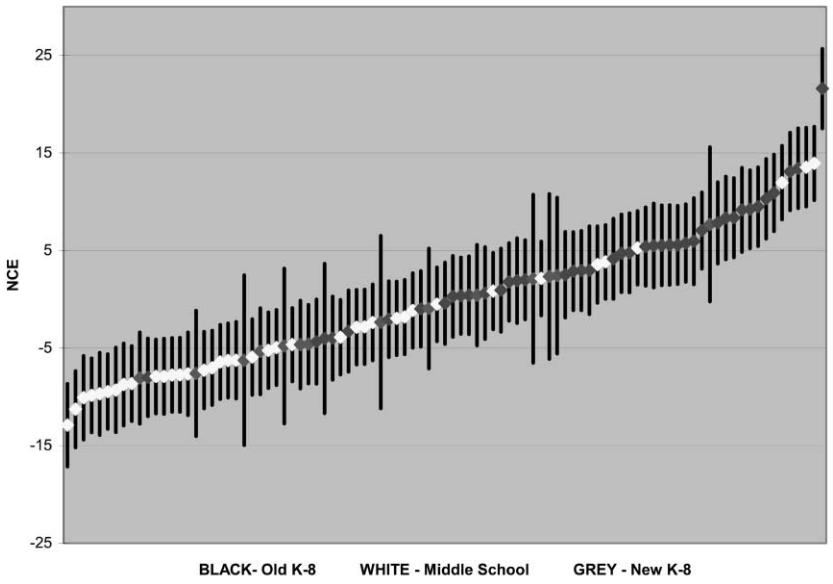

FIG. 1.—Level 3 residuals, empty model. Individual school contributions to mean student achievement (with 95% confidence intervals).

significant advantage in reading but not in math; (4) after controlling for school transition and average grade size, there were no discernible differences between K–8 schools and middle schools in terms of academic achievement.

The differences between K–8 schools and middle schools and the explanatory power of the above-mentioned factors are best seen through the visual representations of figures 1, 2, and 3. Figure 1 shows the level 3 residuals for each school in our data sample, taken from our empty mathematics outcome models prior to the inclusion of any explanatory variables. These residuals can also be thought of as the individual or unique contributions of each school to the average level of math achievement of their student populations (Raudenbush and Willms 1995). In figure 1, the schools are ordered from lowest to highest contribution. Middle schools are highlighted in white, old K–8 schools in black, and new K–8 schools in gray. We are immediately struck by the clustering of old K–8 schools at the top of the figure indicating that as a group they contribute more to the achievement levels of their students than do the set of middle schools. In figure 2, we are shown the same residuals, but after having controlled for time and cohort, prior achievement, and, most important, the external factors of student and teacher demographics. Here we see that old K–8 schools are still clustered toward the top and seem to contribute more as a group, but less so than in figure 1, as there are now

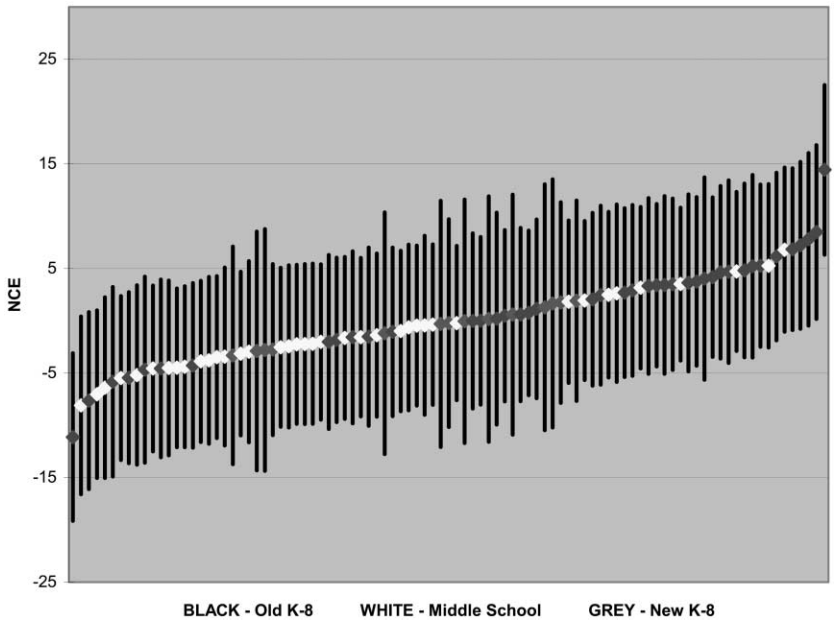

FIG. 2.—Level 3 residuals, middle model. Individual school contributions to mean student achievement (with 95% confidence intervals).

several more old K–8 schools toward the bottom and several more middle schools toward the top. In figure 3, we have the residuals taken from our final model after including all of our significant explanatory variables. Now, after controlling for average grade size and school transitions as well, we find the order is almost reversed and that the old K–8 schools are now clustered toward the bottom.

In all of figures 1, 2, and 3, our set of new K–8 schools did not stand out as a group compared to middle schools, and in each figure they are spread out across the spectrum from high to low, though concentrated more toward the middle. In looking at the descriptive differences between the two school sets earlier in our article, we saw that, while the new K–8 schools enjoyed the traditional K–8 features of smaller size and lower rates of school transitions, in terms of population demographics they were much more like the middle schools in our sample than the older K–8 schools, perhaps even more disadvantaged. In our models, after controlling for student demographics but leaving school transition and grade size uncontrolled, we did see the coefficient for new K–8s in comparison to middle schools rise to about 1.8 and 1.6 NCE in both math and reading, respectively, but not to a level of significance in math, and never to the nominal size or significance level comparable to that

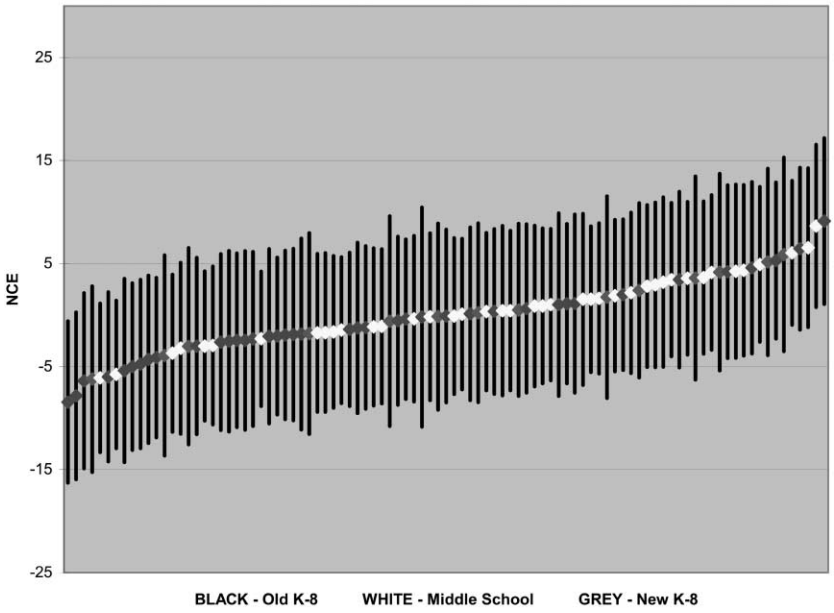

FIG. 3.—Level 3 residuals, full model. Individual school contributions to mean student achievement (with 95% confidence intervals).

of the old K–8 schools in either subject (see table 3). When we then controlled for school transition in the next step, we saw the coefficient for new K–8s decrease by a similar amount as that for old K–8s, by about 0.8 NCE. Then again, when we next controlled for grade size, both new K–8 and old K–8 schools decreased by about 1 NCE in comparison to middle schools across subjects.

As the new K–8 schools did not contribute to mathematics achievement significantly more than middle schools, despite their smaller size and fewer school transitions, we might conclude that these features alone are not enough to replicate the old K–8 school achievement advantage. Even in reading, where the new K–8 advantage became significant after controlling for external factors, it was not as nominally large or as significant as that for old K–8 schools. Much of the old K–8 advantage clearly resides in the different student populations that are served by old K–8 schools and middle schools. We would also believe that the stronger community and relationships that exist in old K–8 schools, which foster student achievement and social outcomes, are not entirely the result of their smallness and continuity into the middle grades but also due to the demographics of their student populations and parents, the

community members themselves. So long as the new K–8 schools consist of the same high-minority and high-poverty student populations as the middle schools, it seems unlikely that they will develop the same sizable achievement advantage seen in the old K–8 schools.

This brings us to an important practical point for any district considering a policy of K–8 conversion. In our sample, between two-thirds to three-quarters of the students were enrolled in middle schools. As these middle schools also have much larger grade sizes on average than do the typical K–8 schools, it would take many more new K–8 schools to replace any set number of middle schools. Given our sample this would be approximately three new K–8 schools for each existing middle school. The other choice would be to create new K–8 schools with much larger average grade sizes, though this would sacrifice what we have shown to be a key to the K–8 achievement advantage.

In the end, even where a district can successfully redistribute its middle school students to K–8 schools of smaller size, we come back to the point that so long as the student demographics remain unchanged, a district is not likely to replicate the K–8 advantage based upon size and school transition alone if its student population remains unchanged. As a policy, then, a district must weigh the infrastructure cost of redistributing middle school students versus the limited achievement gains they may make given the same population demographics.

Where K–8 conversion is not desirable, one solution still open for reformers looking to increase the level of student achievement in middle schools remains the very set of "best practices" that were originally thought to be one of their unique advantages in educating middle-grades students. In Philadelphia many of the highest-performing middle schools, with achievement levels comparable to those of the old K–8 schools, are the ones using outside partner programs designed to implement those best practices such as small learning communities, professional development, cooperative learning, and other pedagogical and classroom instructional strategies (Balfanz et al. 2002).

We must now qualify that, while we have adequately documented the K–8 advantage, our conclusions regarding the new K–8 schools in Philadelphia remain open to debate for two reasons. First, our sample of new K–8 schools included 14 of our 95 schools (15 percent) covering only 32 of our 427 cohorts (8 percent). If the sample of new K–8 schools were larger, and proportionally more equivalent to our sample of middle schools and old K–8 schools, our results, estimates, and tests of significance might all have differed. Second, of the 14 new K–8 schools, seven had added grade 8 for the first time only in the last year of our analysis, 2003–4. Any assessment of the eighth grades at these schools may be premature, and a longer time span should be provided to allow these schools to get past any initial hurdles in expanding the grade

structure and stabilizing their internal environments. Prior research on the implementation of school reforms has found that it can often take three to five years before full and mature implementation is reached and where true gains can be seen and measured (Borman et al. 2003). It may simply take time for these growing and changing schools to build strong community relations.

This uncertainty regarding our estimates for new K–8 schools is reflected in figures 1, 2, and 3, where the smaller number of cases for new K–8 schools leads to inflated standard errors and larger confidence intervals for the estimates of their residuals. Fortunately, as the Philadelphia School District intends to continue with its K–8 conversion policy, we may yet have the chance to follow up on this study several years later with a stronger evaluation of achievement levels in new K–8 schools.

One final practical note regarding the K–8 advantage and the K–8 conversion policy is for reformers to consider the actual size of the K–8 advantage, approximately 3.2 NCE in math and 1.8 in reading after controlling for external population factors. In Philadelphia, one of the main targets for making adequate yearly progress is the percentage of students scoring below "Basic" on the PSSA, a standard that is equivalent to a scale score of 1,180 in both subjects. In our sample, 60 percent of the middle school students were below this mark in math, and the NCE mark at which they were below was the thirty-seventh NCE. The 3.2 NCE bump in average student achievement that they would have hypothetically received by attending K–8 schools instead would bring up all the students from the thirty-third NCE and above to the Basic performance level. This would (in theory) have led to a reduction in the percent of below-Basic students by roughly 7 percent, to 53 percent of all middle school students. In reading, 56 percent of middle school students scored below Basic, and with the hypothetical bump of 1.8 NCE they would have received had they instead attended K–8 schools, another 5 percent of them might have been above Basic, reducing the total to 51 percent scoring below Basic.

This would indeed be a sizable reduction in the percent of students scoring below Basic, but it still leaves over 50 percent of students scoring below in both subjects. These impacts are also based upon the old K–8 schools, and the new K–8 schools serving more disadvantaged populations did not perform as high. The 3.2 NCE old K–8 advantage translates into an effect size of 0.19 in terms of mathematics achievement, and the reading advantage of 1.8 equals an effect size of 0.11. Even if such gains were realized through conversion, they may not be enough to close the achievement gap that exists for minority and high-poverty students in the United States or the gap between the U.S. public schools and other international countries' middle-grades students. A K–8 conversion policy alone does not represent a "silver bullet" reform for

closing the achievement gap and improving student achievement, and administrators must ask themselves if such a massive reform is truly worth the resources given the likely impacts. They must also compare it to other possible reforms and decide if they are getting with K–8 conversions the best possible "bang for their buck" in terms of reform finances.

Moving beyond the K–8 reform and to reform policies in general, our study has revealed some other relevant lessons for policy makers. Prior to including any of our control variables, even those for K–8 schools, we had already come across an important result. Three-quarters of the variation in achievement was at the student level (76 percent; $p < .001$), and separate from variation between schools and cohorts. As government institutions and school districts continue to push schools to improve their achievement performances, they must also ask to what degree schools and school-based reforms will be able to effect student achievement. If the majority of variation in achievement pertains to the students themselves, the current ideas regarding school reforms and the linking of school's annual performances to reward and punishment systems might be the wrong methods for reaching the right goals. Even with all of our explanatory measures, our models still explained less than half of the between-student variation in achievement. It is likely that a good deal of that unexplained variation resides in factors pertaining to a student's parents and their home environment, factors that schools and school administrators cannot address on a schoolwide level.

Furthermore, many of the new accountability systems that have been put in place since the No Child Left Behind act rely on measuring schools' yearly performances on standardized tests such as the PSSA. Schools can face severe punishments ranging from a reduction of funding to staff restructuring and the dismissal of administrators, all based on yearly changes in their mean test scores. If, however, as we have seen here, there is a significant amount of random variation in achievement between cohorts themselves (6 percent; $p < .001$), a year-to-year dip in mean test scores may not be reflective of school, staff, or administrator performance but rather an indicator of the difference between two unique cohorts of students. Many parents with more than one child can talk of one of their children as having an academically strong cohort and another sibling as having an average or poor cohort in comparison, when thinking of their children's friends and classmates in their grade.

In conclusion, we have found that K–8 schools do on average have higher levels of achievement. This advantage is due partially to differences in the populations of these schools and partially to structural differences. In the end, the advantage is multifaceted and not easily replicated. Districts and schools eager to convert to the K–8 structure because of this advantage should not rush into any such policies but rather should reflect upon history. K–8 schools, once the dominant school structure in the U.S. middle-grades landscape, have

fallen out of fashion before, and they may yet do so again as the rush to revert to them is likely to leave many reformers disappointed.

References

Balfanz, Robert, Kurt Spiridakis, and Ruth Neild. 2002. "Will Converting High-Poverty Middle Schools to K–8 Schools Facilitate Achievement Gains?" Report by the Philadelphia Education Fund, Philadelphia.

Beaton, Albert E., Ina V. S. Mullis, Michael O. Martin, Eugenio J. Gonzalez, Dana L. Kelly, and Teresa A. Smith. 1996. "Mathematics Achievement in the Middle School Years." TIMSS Study Center, Boston.

Borman, Geoffrey, Gina M. Hewes, and Laura T. Overman. 2003. "Comprehensive School Reforms and Achievement: A Meta-analysis." *Review of Educational Research* 73 (2): 125–230.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models*. Newbury Park, CA: Sage.

Burrill, Gail. 1998. "Changes in Your Classroom: From the Past to the Present to the Future." *Mathematics Teaching in the Middle School* 4:184–90.

Coladarci, Theodore, and Julie Hancock. 2002. "The (Limited) Evidence Regarding Effects of Grade-Span Configurations on Academic Achievement: What Rural Educators Should Know." *Journal of Research in Rural Education* 17 (3): 189–92.

Eccles, Jacqueline S., Sarah Lord, and Carol Midgley. 1991. "What Are We Doing to Early Adolescents? The Impact of Educational Contexts on Early Adolescents." *American Journal of Education* 99 (4): 521–42.

Eccles, Jacquelynne S., and Carol Midgley. 1989. "Stage/Environment Fit: Developmentally Appropriate Classrooms for Early Adolescents." In *Research on Motivation in Education*, vol. 3, ed. Russel E. Ames and Carol Ames. New York: Academic Press.

Epstein, Joyce L., and Douglas J. MacIver. 1990. *Education in the Middle Grades: Overview of National Practices and Trends*. Columbus, OH: National Middle School Association.

Herman, Barry E. 2004. "The Revival of K–8 Schools." *Phi Delta Kappa Fastbacks* 519: 7–37.

Hough, David L. 2005. "The Rise of the 'Elemiddle' School." *School Administrator* 62 (3): 10–14.

Jackson, Anthony W., and Gayle A. Davis. 2000. *Turning Points 2000: Educating Adolescents in the 21st Century*. New York: Teachers College Press.

Kao, Grace. 1995. "Asian Americans as Model Minorities? A Look at Their Academic Performance." *American Journal of Education* 103 (2): 121–59.

Lee, Valerie E., and Julia B. Smith. 1993. "Effects of School Restructuring on the Achievement and Engagement of Middle Grade Students." *Sociology of Education* 66 (3): 164–87.

McEwin, C. Kenneth, and Thomas S. Dickinson. 1996. "Middle-Level Teacher Preparation and Licensure." In *What Research Says to the Middle-Level Practitioner*, ed. Judith L. Irvin. Columbus, OH: National Middle School Association.

McEwin, C. Kenneth, Thomas S. Dickinson, and Doris M. Jenkins. 1996. *America's Middle Schools: Practices and Progress—a 25-Year Perspective*. Columbus, OH: National Middle School Association.

McEwin, C. Kenneth, Thomas S. Dickinson, and Michael G. Jacobson. 2005. "How Effectice Are K–8 Schools for Young Adolescents?" *Middle School Journal* 37 (1): 24–28.

Midgley, Carol. 1993. "Motivation and Middle-Level Schools." In *Motivation and Adolescent Development*. Vol. 8 of *Advances in Motivation and Achievement*, ed. Martin L. Maehr and Paul R. Pintrich. Greenwich, CT: JAI.

Mizell, Hayes. 2005. "Grade Configurations for Educating Young Adolescents Are Still Crazy after All These Years." *Middle School Journal* 37 (1): 14–23.

National Forum to Accelerate Middle Grades Reform. 2002. *Policy Statement: Teacher Preparation, Licensure, and Recruitment*. Newton, MA: Author.

Offenberg, Robert. 2001. "The Efficacy of Philadelphia's Schools Compared to Middle Grades Schools." *Middle School Journal* 32 (4): 23–29.

Paglin, Catherine, and Jennifer Fager. 1997. *Grade Configuration: Who Goes Where?* Portland, OR: Northwest Regional Educational Laboratory.

Pardini, Priscilla. 2002. "Revival of the K–8 School." *School Administrator* 59 (3): 6–12.

Peng, Samuel S., and DeeAnn Wright. 1994. "Explanation of Academic Achievement of Asian American Students." *Journal of Educational Research* 87 (July/August): 346–52.

Raudenbush, Stephen W., and J. Douglas Willms. 1995. "The Estimation of School Effects." *Journal of Educational and Behavioural Statistics* 20 (4): 307–35.

Reising, Bob. 2002. "Middle School Models." *The Clearing House* 76 (2): 60–61.

Schmidt, William H., Curtis C. McKnight, Pamela M. Jakwerth, Leland S. Cogan, Richard T. Houang, et al. 1999. *Facing the Consequences: Using TIMSS for a Closer Look at the United States Mathematics and Science Education*. Dordrecht: Kluwer.

Simmons, Robert, Ann Black, and Yingzhi Zhou. 1991. "African-American versus White Children and the Transition into Junior High School." *American Journal of Education* 99 (4): 521–42.

Simmons, Robert, and Dale Blyth. 1987. *Moving into Adolescence: The Impact of Pubertal Changes and School Context*. New York: Aldine de Gruyter.

Snijders, Tom A. B., and Roel J. Bosker. 1999. *Multilevel Analysis*. Thousand Oaks, CA: Sage.

Weiss, Christopher, and Lindsay Kipnes. 2006. "Reexamining Middle School Effects: A Comparison of Middle Grades and K–8 Schools." *American Journal of Education* 112 (2): 239–72.

Yakimowski, Mary E., and Faith Connolly. 2001. *An Examination of K–5, 6–8, and K–8 Grade Configurations*. Report prepared for the Board of School Commissioners. Baltimore: Division of Research, Evaluation, and Accountability, Baltimore City Public School System.