"Comparing Achievement between K-8 & Middle Schools: A Large Scale Empirical Study"

Vaughan Byrnes & Allen Ruby
Center for Social Organization of Schools
Johns Hopkins University

The research reported here was supported by the Research on Learning and Education (ROLE) Program at the National Science Foundation, grant number 0411796

ABSTRACT:

This paper reports on a natural experiment in the Philadelphia School District. It compares Middle Schools to established K-8 schools, as well as to newly formed K-8 schools that are part of the district's K-8 conversion policy, in order to determine if the different school structures have an effect on student academic achievement. The outcome is students' 8th grade mathematics achievement and our sample includes 40,883 8th grade students taken from 95 schools across 5 cohorts from the 1999-2000 to the 2003-04 school years. The analysis uses multi-level modeling to account for differences between student, cohort, and school level variation, and includes a large set of statistical controls that include student demographics, teacher characteristics, school transition, and several cohort and school level factors including average school size.

The results find that older K-8 schools do perform significantly better than Middle Schools, and that this advantage is adequately explained by the two school type's differing student and teacher populations, differences in average grade size, and the extra school transition that Middle School students must make from elementary to the middle grades. Newer K-8 schools created as part of the district's reform efforts outperformed Middle Schools but not by as much or as significantly as did older K-8 schools, despite having smaller grade sizes and lower rates of school transition. We found that this was likely due to their student populations, which like those of Middle Schools, consisted primarily of minority students from high-poverty backgrounds. We conclude that while K-8 schools do perform better in terms of student achievement, the advantage exists for several reasons and may not be easily replicated or represent a solution to the problem of low achieving schools and students in large urban public school districts that serve high-minority and low-poverty student populations.

This paper reports on a natural experiment in the Philadelphia City School District. Using longitudinal data and multi-level modeling, we compare students who attended Middle Schools (middle grades only) to students from K-8 schools (elementary through 8th grade) 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 amongst 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 in the U.S. has widened, along with the gap between the U.S. as a whole and the comparative levels of achievement in other developed nations, many large scale and high-resource reform efforts have been undertaken over the last decade with the direct aims of improving student achievement in these schools (Burrill, 1998). One of the more popular reforms currently 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 at some new innovation, the conversion to the K-8 model represents in some way a return to the old.

It was the K-8 school that predominated middle grades education in the United States at the end of the 19th century and through the first four decades of the 1900's, only to be supplanted by the Junior High model (7th to 9th) that began in the 1910's and became the predominant school structure by the 1960's. The Junior High was then itself replaced by the Middle School model, which rose in the 1960's and 1970's to become the dominant school structure of the 1990's (Mizell, 2005; Herman, 2004; Paglin & 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; Reising, 2002; Pardini, 2002). In some ways the K-8 model has also never left, as it has remained a popular choice amongst 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 grades 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 behavioural needs specific to 10-13 year olds (Coladarci et. al., 2002; Yakimowski & 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 semi-departmentalization), and mixed level classrooms – that would allow them unique advantages in addressing the academic achievement of middle grades

students (Hough, 2005; Offenberg, 2001; Midgley, 1993; Lee & Smith, 1993; Epstein & MacIver, 1990). 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 non-academic 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 et. al., 2002; Offenberg, 2001; Yakimowski & Connolly, 2001). As academic achievement in terms of performance on standardized tests is the fundamental method of evaluation for both districts and states, and their main measure for 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 (Pardini, 2002; Coladarci et. al., 2002) and better performance in terms of emotional and social outcomes such as self-esteem, leadership, and attitudes towards school (Weiss & Kipnes, 2006; Simmons & Blyth, 1987). 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 & 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 neighbourhood school adds to their greater sense of community (Mizell, 2005; Herman, 2004). 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 a K-5 or K-6 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 & Connolly, 2001). Another reasons for districts to convert is that K-8 schools are often more cost efficient in terms of building and property maintenance (one building versus 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 5th or 6th grade; 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 & 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 neighbourhoods, leading to poorer performance in school and limits on their lifelong opportunities. Such demographic factors may 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 regardless of the causes of these demographic differences, converting Middle Schools into K-8s and a policy change in grade span alone may not lead to a significant improvement in student achievement given similar student demographics.

Teacher Population

Similarly, teacher characteristics such as years of experience, levels of certification, retention rates, and student-teacher ratios, are also thought to be differences between K-8 and Middle Schools that contribute to their diverging performances on student achievement and social outcomes (Paglin & Fager, 1997; Simmons & Blyth, 1987). As most teachers are trained to teach at either the elementary or high school levels, many Middle Schools 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 (National Forum to Accelerate Middle Grades Reform, 2002; Jackson & Davis, 2000; McEwin & Dickinson, 1996; McEwin, Dickinson, & Jenkins, 1996). 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 aged students and the particular set of teaching practices that were recommended for teachers of middle grades and may require the 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 a new team of staff, a

change that can be harmful for student performance and engagement, especially amongst minority students (Herman, 2004; Coladarci, 2002; Simmons et. al., 1991; Simmons & Blyth, 1987).

An offshoot of this transition is that when Middle School students make the switch to a new middle grades 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 among others (Herman, 2004; Coladarci, 2002; Yakimowski & Connolly, 2001; Simmons & Blyth, 1987). 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 it 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 (Weiss & Kipnes, 2006; Coladarci, 2002; Offenberg, 2001; Lee & Smith, 1993; Eccles et. al., 1991; Simmons & Blyth, 1987). 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 should not be wandering the hallways. Thus, smaller size is also a probable a cause of the

stronger sense of community visible in K-8 schools, as smaller size allows for students, teachers, and parents to foster closer and stronger relationships (Offenberg, 2001; Paglin & 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 for establishing mixed level classrooms (Hough, 2005; Coladarci, 2002; Yakimowski & Connolly, 2001; Offenberg, 2001; Eccles & Midgley, 1989).

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 oppose 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 (Weiss & Kipnes, 2006; Hough, 2005; McEwin et. al., 2005; Balfanz, 2002; Pardini, 2002; Coladarci, 2002; Yakimowski & 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 Weiss & Kipnes, Offenberg, and Simmons & Blyth employed rigorous statistical analyses. Weiss & Kipnes (2006) used a comparative sample and multi-level modeling and found that students at K-8 schools reported higher levels of self-esteem and 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 & Blyth (1987) was a student level analysis that found that students from K-8 schools enjoyed high levels of social engagement, better attitudes towards school, and also higher levels of self-esteem.

In this study we take a step forward and merge the two techniques by using multi-level modeling for our analyses (Snijders & Bosker, 1999; Bryk & Raudenbush, 1992). Multi-level models, or hierarchical linear models, account for grouped data such as ours, where students are nested within schools. Multi-level 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 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 is 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 2003-04 year and covers 40,883 8th grade students, taken from 95 schools over the 5 cohorts (the first graduated from 8th grade in the 1999-00 school year, and the last in 03-04). 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 N (between school), 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 5 years observed in our study 14 elementary schools were transformed into new K-8 schools and

so did not have 8th grades throughout all the years of analysis. These elementary schools added one grade per year, and of these 14 new K-8 schools, 1 reached 8th grade in spring '00, 4 in spring '01, 2 in '03, and 7 added the 8th grade in spring '04. Conversely 6 Middle Schools were being transformed into upper grades schools ranging from 5-12 or 6-12 schools, junior highs or high schools. Their transformations began in the 03-04 year at which point their cohorts in that year were excluded, as the schools no longer fit into the Middle School versus K-8 comparison. Missing data also cost us the inclusion of 1 cohort in one school, and finally one Middle School ceased to exist entirely as part of the school district after the 2000-01 school year.

Of our K-8 schools, 1 was actually a grades 1st-8th school, and of our Middle Schools, 17 were 5th-8th, 20 were 6th-8th,and two were originally 6th-8th but transformed into 7th-8th during the period of observation. No K-12 schools were examined as they fall 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. 5 district 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 neighbourhood 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.

This represented the entire population of schools from the Philadelphia School District, as well as the population of cohorts to pass through these schools during the

period of observation. However, our study includes only a sample of the student population as missing data led to the exclusion of a proportion of students, discussed in more detail later in the paper. In the end, we possessed a very large sample of students from all schools in the PSD over all the years examined, as we tried to assess and explain any differences between the two school structures and evaluate the early returns on the K-8 conversion reforms of the Philadelphia School District.

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 only been able to focus on a select few each. 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 was 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' 8th 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 1st and 2nd NCE equivalent to the

distance between the 49th and 50th, unlike with percentiles. NCE 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 8th grade PSSA as an outcome is highly appropriate as it is one of the 'high-stakes' tests 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 too, especially since the introduction of the 'No Child Left Behind' 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' 5th grade scores on the PSSA, used to control for their prior levels of achievement (5th grade is the last year in which the test is administered prior to the 8th 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 5th grade, as several Middle Schools begin in the 5th 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 (Snijders & Bosker, 1999; Bryk & Raudenbush, 1992). 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. Students' prior scores were then also aggregated to the cohort level, to see if the mean prior achievement level of their cohorts affected their individual scores in 8th grade.

Time

In exploratory analyses, we examined the use of different measures to account for both change over time in student test scores and changes between cohorts. Our exploratory models included 8 cohorts, whereas our final models include only 5. This is because while our data runs from the 1996-97 school year to 2003-04, that eight-year span only includes the prior 5th grade scores for the last five cohorts that graduated 8th grade from '00 to '04, and finished 5th grade between '97 and '01. However, when analyzing the longer time span of eight years in exploratory models, 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 8th and 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 upwards.

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 which 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 is not released for individual students and thus can only be included 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 & 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, '99-'00. However, exploratory analyses were run with the sub-sample of our data for which teacher retention was available and they did not find any significant correlations between the percent of returning teachers and students' achievement scores, 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 8th grade as they had been in the 4th 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 8th grade as in 4th, 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 8th grade year, and another measure that accounts for school level mobility rates during the school-year (measured by summing the number of students who both entered the school after the start

of the year and the number that left before it ended, divided by the total enrolment of the school during that year). 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 8th grade alone, the average size of the middle grades (5th through 8th), 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 the middle grades of a school alone, 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 higher SES areas (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 old K-8 schools, and newly formed K-8 schools (created in the last 5 years as part of the district's reform policy), both 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 it conversion policy.

Student Attendance

Finally, student attendance, both at the individual student level and aggregated to the cohort level were originally included in our exploratory analyses. However, they were excluded for the following reasons. First, student attendance at the cohort level was highly correlated to the measures for % of FRL students (-0.573; $P<0.001$), % of minority students (-0.443; $P<0.001$), and average grade size (-0.542; $P<0.001$), thus creating problems with modeling, and preventing the inclusion of all the variables in the models in unison due to collinearity. However, while student attendance was strongly correlated to all three of the other measures, the other three measures were much less correlated to each other. Further reasoning is that while each of the other three measures, % FRL, % minority, and average grade size could have causally led to different rates of student attendance, the inverse cannot be true, and student attendance could not logically be a causal precedent of the other three measures. Also relevant, is that measures of student

attendance, while obviously a strong predictor of student achievement, are also considered to be an advantage of K-8 schools compared to Middle Schools, and thus might be inappropriate for inclusion in our models as a statistical control where they might serve to mask or lessen a possible K-8 advantage in academic achievement.

Descriptive Comparison

Tables 1-2 show the results from T-tests comparing the different school structures along the above mentioned-variables. The tables, and variables, are divided into three sets, one each for the student, cohort, and school level factors. 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 5^{th} and 8^{th} 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 8^{th} grade as in 4^{th}, 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 3 years greater experience compared to teachers at Middle Schools,

while also having lower rates of teacher absenteeism, and greater proportions of certified teachers. All the above differences between old K-8 schools and Middle Schools were significant at the α -.001 level, confirming that K-8 schools did indeed have higher levels of student achievement, but at the same time served significantly different 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 students, 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 that 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 2003-04 and final cohort, while the number of students attending K-8 schools increases. In 99-00, 73% of that cohort's students were enrolled in Middle Schools and 27% in K-8 schools. By 03-04, 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% of students enrolled in Middle School versus 36% at K-8 schools as we see the district's K-8 conversion policy taking effect.

Hypotheses

Moving forward to our multi-level model analysis, the question of comparing Middle Schools to K-8 schools has become more intricate. Both the old K-8 and 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 school in fact have worse off 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 different 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 multi-level models, we chose to build ours in a series of three 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 three steps begin with an empty model, and then add in measures for time, cohort, and prior achievement, followed by population demographics, and then 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 & 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 backwards fashion starting with the least significant estimates. This kept with the idea of valuing parsimony among other things, as multi-level 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 in various different ways. 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 α-.001 level for all three levels of our model, given the χ^2 statistics provided by HLM software Bryk & 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% of the

variation in math achievement is between students at Level 1 of our model, 6% at Level 2 between cohorts, and that the remaining 18% varies between schools at Level 3. For reading, the numbers were 79%, 4%, and 15%. This is consistent with other research on school effects that has found between 10-30% of the variability in student achievement is typically between schools (Bryk & 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 - β=8.55, t*=5.89, P<0.000; Read - β=8.23, t*=5.70, P<0.000). 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 8th and final cohort of 03-04 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 5th grade score, was then added to our model at both the student level and also 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 - β=3.60, t*=5.12, P<0.000; Read - β=3.97, t*=6.24, P<0.000). However, this merits some qualification as it does not so much halve the effect of K-8 schools, but rather halves 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 the 5th grade. Controlling for their achievement at 5th grade, 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 8th 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% 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 scores 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% 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 - β=3.23, t*=4.31, P<0.000; Read - β=2.34, t*=4.11, P<0.000), 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 impacted by the additions of teacher quality controls (β=3.18, t*=4.13, P<0.000), there was a substantial impact on the reading achievement advantage (β=1.75, t*=3.23, P<0.002). Controlling for teacher characteristics also brought the difference between new K-8 and Middle Schools to a level of significance in reading, (β=1.61, t*=2.07, P<0.041), 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 (β=1.68, t*=2.17, P<0.033), while pushing its statistical significance above the α-.010 level. In reading, the

old K-8 school advantage was also reduced by 1 NCE (β=0.65, t*=1.09, P<0.278) 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 (β=0.52, t*=0.63, P<0.531) and making it non-significant. Overall, students who were in the same school in 4th grade as in 8th scored on average almost 2 NCE higher in both math and reading compared to 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 (β=0.15, t*=0.15, P<0.881), and further reducing the old K-8 advantage in reading (β=0.54, t*=0.57, P<0.567). The differences between new K-8 schools and Middle Schools, was 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 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. Compared to the Southern region, where the plurality of schools is based, the Western region was much worse and the Northwestern region performed much better. 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% of the total variation in mathematics achievement was explained by the explanatory variables in our final model. 45% of the variation between students was explained, along with 65% of the cohort level variation, and 44% of the variation between schools. In reading, our model led to a proportional reduction of 46% in the total variation for reading achievement, 44% of the between student variation, 63% of the cohort variation, and 53% of the variation between schools. Comparing the deviance statistics (-2logLikelihood) of our empty and final models, we arrive at χ^2 test statistics of 23,947.2 with 55 degrees of freedom for math, and 23064.9 with 62 DF for reading. Both result in highly significant P-values at the α-.001 level, confirming that the explanatory variables

included in our final models have in fact contributed statistically significant information regarding our outcome variables, students' 8th grade achievement scores.

Strengths & Limitations

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.

Statistical Conclusions

That student achievement varies randomly at both the cohort and school level represents empirical evidence that multi-level models are in fact an appropriate and necessary method for analyzing the research questions presented in this paper. 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 neighbourhood 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 8th 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 on each student, or case, that we considered. 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 8th grade score, and all their demographic data, but lacked their 5th grade score for prior achievement or 4th 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% to a low of 24%. However, only 8% (34 of the 428 cohorts) were below 50% representation, and conversely 92% of cohorts were represented by half of their students or more. 68% of the cohorts were over 60% representation, 28% over 70%, 7% over 80%, and 3% over 90. The representation of students was slightly higher in K-8 schools where attendance is higher, where on average 67% of the students from each cohort were included compared to only 60% for Middle Schools. Furthermore, we were able to compare the true cohort averages for our outcomes, 8th grade PSSA scores, to our sample averages, though only in terms of Scale Scores, and not the NCE metric used in the study. Knowing 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 1183 in math and in 1168

reading, our sample means were 1197 in math and 1192 in reading, thus overestimates 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% 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 8th grade student back to their 5th grade prior score and their 4th grade prior school as they fall out of their original cohort. In addition, the nature of 3-Level multi-level models, which require each Level-1 unit (students) to be hierarchically nested in only 1 Level-2 unit (cohorts), also presents a problem as students who repeat a grade would them 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 (Weiss & Kipnes, 2006; Simmons & Blyth, 1987) 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 regards to these factors at least, our study provides any conclusive results. As all our teacher measures were aggregated at 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 8th 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 8th 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 the 5th grade 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 & Blyth (1987). Since no Middle School students were in the same school in 4th and 8th grades (by definition), and since no K-8 students switched to a MS for the middle grades (or they would be classified as Middle School in our model), then we 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. 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 4th to 8th grade, 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 paper 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 are external to the school structures such as student and teacher demographics, two other reasons are smaller size and the greater continuity that are largely intrinsic to K-8s and synonymous with the reform.

External & 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 that we have used, the large representative samples of schools, cohorts, and students that our analysis incorporates over a wide time frame, and the large set of theoretically relevant statistical controls that we have been able to incorporate.

Beyond this study, and in regards 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 schools 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 different 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 significant advantage in reading but not in math; 4) after controlling for school transition and average grade size, there were no discernable 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

Charts 1, 2, & 3. Chart 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 & Willms, 1995). In Chart 1 the schools are ordered from lowest to highest contribution. Middle Schools are highlighted in yellow, old K-8 schools in blue, and new K-8 schools in brown. We are immediately struck by the clustering of old K-8 schools at the top of the chart indicating that as a group they contribute more to the achievement levels of their students than do the set of Middle Schools. In Chart 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 towards the top and seem to contribute more as a group, but less so than in Chart 1 as there are now several more old K-8 schools towards the bottom and several more Middle Schools towards the top. In Charts 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 towards the bottom.

In all of Charts 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 chart they are spread out across the spectrum from high to low, though concentrated more towards the middle. In looking at the descriptive differences between the two school sets earlier in our paper, 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 to 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-8's in comparison to Middle Schools rise to about 1.8 to 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 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-8's decrease by a similar amount as that for old K-8's, by about 0.8 NCE. Then again when we next controlled for grade size, both new K-8 and old K-8 schools decrease 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 sizeable 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 3 new K-8 schools for each existing Middle School. The other choice would be 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 is predominantly from high-minority and high-poverty backgrounds. 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’ which were originally thought to be one of their unique advantages in educating middle grades students. While many Middle Schools may be doing a poor job of implementing these practices, and many K-8 schools might be using them more frequently with greater ease (Hough, 2005), 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%) covering only 32 of our 427 cohorts (8%). 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. Secondly, of the 14 new K-8 schools, 7 had only added 8th grade for the first time in the last year of our analysis, 03-04. Any assessment of the 8th 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 3-5 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 Charts 1, 2, & 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, possibly to the point of eliminating all its Middle Schools in favour of K-8 schools, 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 AYP is the percentage of students scoring below 'Basic' on the PSSA, a standard which is equivalent to a scale score of 1180 in both subjects. In our sample, 55% of the students were below this mark in math, 60% in Middle Schools, and the NCE mark at which they were below was the 37th NCE. For Middle Schools students, 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 33rd 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%, to 53% of students. In reading, 56% 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% of them might have been above basic, reducing the total to 51% scoring below basic.

This would indeed be a sizeable reduction in the percent of students scoring below basic, but it still leaves over 50% 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 with K-8 conversions, they are getting 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%; $P<0.001$), 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 school wide level.

Furthermore, many of the new accountability systems that have been put in place since 'No Child Left Behind' 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 tests scores. If however, as we have seen here, there is a significant variation in achievement between cohorts themselves (6%; $P<0.001$), 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 multi-faceted 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.

Table 1
Student Level Descriptives

Variable	Middle School	Old K-8	New K-8
N =	**28,595**	**10,938**	**1,350**
8th Grade Math Score	34.0 (16.6)	41.9*** (17.0)	33.1* (14.7)
5th Grade Math Score	29.9 (17.3)	35.4*** (17.4)	24.7*** (14.9)
8th Grade Reading Score	34.5 (16.6)	42.1*** (17.4)	33.9 (14.3)
5th Grade Reading Score	30.2 (16.8)	35.4*** (17.8)	26.4*** (15.0)
Female	53%	53%	53%
Special Educ.	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 4th & 8th	0%	63%***	63%***

** - significant at .05 level*
*** - significant at .01 level*
**** - significant at .001 level*
(parenthesis) – standard deviations

Table 2
Cohort Level Descriptives

Variable	Middle School	Old K-8	New K-8
N =	**187**	**208**	**32**
New Principal	25%	21%	34%
Free/Reduced Lunch Program	79.7% (16.6)	68.0%*** (21.5)	93.3%*** (5.1)
Minority (Hispanic + Black)	85.6% (20.1)	64.8%*** (24.8)	96.3%*** (9.7)
Average Grade Size	248 (85.0)	74*** (30.4)	76*** (22.5)
Student Mobility	37.5% (11.5)	30.9%*** (11.9)	44.2%** (7.8)
Mean 5th Grade Math Score	29.2 (7.6)	34.9*** (7.6)	24.9** (6.1)
Mean 5th Grade Reading Score	29.7 (6.8)	34.8*** (7.3)	26.5*** (4.2)
Teacher Absentee Rate	6.6% (2.0)	5.7%*** (2.1)	6.5% (1.9)
Certified Teachers	85%	93%***	81%*
Teacher Experience	11.7 (3.3)	14.8*** (4.4)	9.5*** (2.6)
Student/Teacher Ratio	18.1 (2.5)	17.6* (2.6)	17.3 (2.2)
Cohort '00	21%	20%	3%***
Cohort '01	21%	20%	16%
Cohort '02	21%	20%	16%
Cohort '03	20%	20%	22%
Cohort '04	17%	20%	44%**

** - significant at .05 level*
*** - significant at .01 level*
**** - significant at .001 level*
(parenthesis) – standard deviations

Table 3
Coefficient Estimates for Old K-8 and New K-8 Schools
(compared to Middle Schools)

	MATHEMATICS				READING			
	Old K-8							
	β	**t***	**P-Value**	**95% CI**	**β**	**t***	**P-Value**	**95% CI**
With no Statistical Controls	8.56 (1.45)	5.89	0.000***	11.4 5.7	8.23 (1.37)	6.01	0.000***	10.9 5.6
After Including measures for Time & Cohort	8.41 (1.42)	5.93	0.000***	11.2 5.6	7.97 (1.41)	5.64	0.000***	10.7 5.2
After controlling for Prior Achievement	3.60 (0.70)	5.12	0.000***	5.0 2.2	3.97 (0.64)	6.24	0.000***	5.2 2.7
After adding Student Demographics	3.23 (0.75)	4.31	0.000***	4.7 1.8	2.34 (0.57)	4.11	0.000***	3.5 1.2
After adding Teacher Characteristics	3.18 (0.77)	4.13	0.000***	4.7 1.7	1.75 (0.54)	3.23	0.002**	2.8 0.7
After Including School Transition	1.68 (0.77)	2.17	0.033*	3.2 0.2	0.65 (0.59)	1.09	0.278	1.8 -0.5
After controlling for Average Grade Size	0.15 (0.97)	0.15	0.881	2.1 -1.8	0.54 (0.95)	0.57	0.567	2.4 -1.3
	New K-8							
	β	**t***	**P-Value**	**95% CI**	**β**	**t***	**P-Value**	**95% CI**
With no Statistical Controls	1.16 (1.84)	0.63	0.529	4.8 -2.5	0.67 (2.06)	0.326	0.745	4.7 -3.4
After Including measures for Time & Cohort	-1.40 (1.74)	-0.80	0.425	2.0 -4.8	-1.41 (1.39)	-1.02	0.312	1.3 -3.9
After controlling for Prior Achievement	1.06 (1.16)	0.92	0.362	3.3 -1.2	0.94 (0.90)	1.04	0.301	2.7 -0.8
After adding Student Demographics	1.68 (1.15)	1.46	0.144	3.9 -0.6	1.49 (0.82)	1.81	0.073	3.1 -0.1
After adding Teacher Characteristics	1.79 (1.15)	1.56	0.122	4.0 -0.5	1.61 (0.78)	2.07	0.041*	3.1 0.1
After Including School Transition	0.90 (0.92)	0.98	0.330	2.7 -0.9	0.52 (0.82)	0.63	0.531	2.1 -1.1
After controlling for Average Grade Size	-0.24 (1.05)	-0.23	0.819	1.8 -2.3	0.49 (1.03)	0.40	0.632	2.5 -1.5

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	0.000***	36.31	2.37	0.000***
% FRL, G010	-0.04	0.02	0.019*	-0.03	0.01	0.029*
Mobility, G020	-5.56	2.31	0.018*	-5.58	2.20	0.012*
S/T Ratio, G030	-0.13	0.08	0.119	-0.19	0.08	0.022*
Experience, G040	0.04	0.09	0.636	0.22	0.07	0.003**
Grade Size, G050	-0.008	0.003	0.004**	-0.003	0.002	0.059
Time, G060	0.99	0.17	0.000***	0.62	0.14	0.000***
Cohort 99-00, G070	1.64	0.45	0.001***	1.29	0.37	0.001***
Slope for Prior Achievement, 5th Grade NCE						
Intercept, G100	0.60	0.01	0.000***	0.60	0.01	0.000***
Slope for Gender						
Intercept, G200	0.59	0.14	0.000***	2.31	0.13	0.000***
Slope for Special Education						
Intercept, G300	-0.42	0.28	0.132	-1.02	0.32	0.002**
Slope for ESL						
Intercept, G400	-0.98	0.47	0.038*	-2.13	0.40	0.000***
Slope for White						
Intercept, G500	1.11	0.24	0.000***	1.09	0.27	0.000***
Slope for Asian						
Intercept, G600	7.17	0.34	0.000***	5.63	0.37	0.000***
Slope for Hispanic						
Intercept, G700	0.73	0.19	0.000***	0.73	0.20	0.000***
Slope for Same School in 5th & 8th Grades						
Intercept, G800	1.61	0.32	0.000***	1.94	0.25	0.000***

Chart 1
Level 3 Residuals, Empty Model
Individual School Contributions to Mean Student Achievement
(with 95% Confidence Intervals)

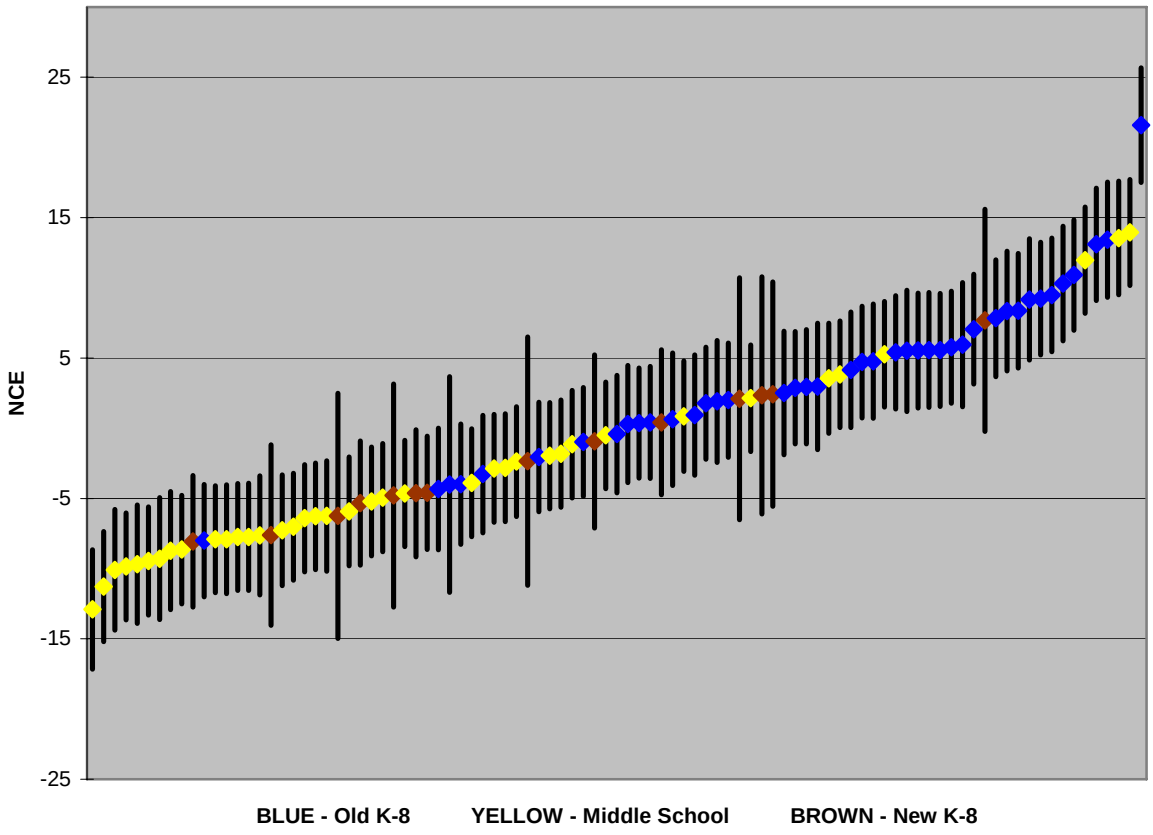


Chart 2
Level 3 Residuals, Middle Model
Individual School Contributions to Mean Student Achievement
(with 95% Confidence Intervals)

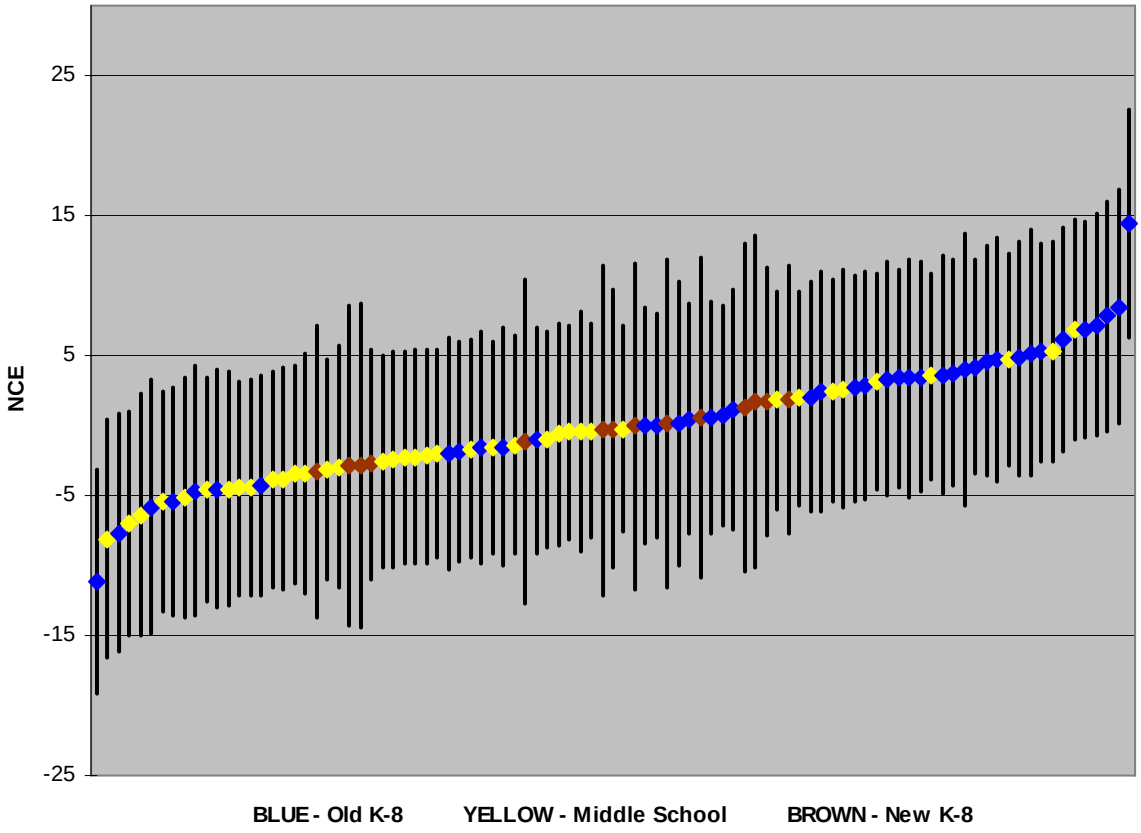
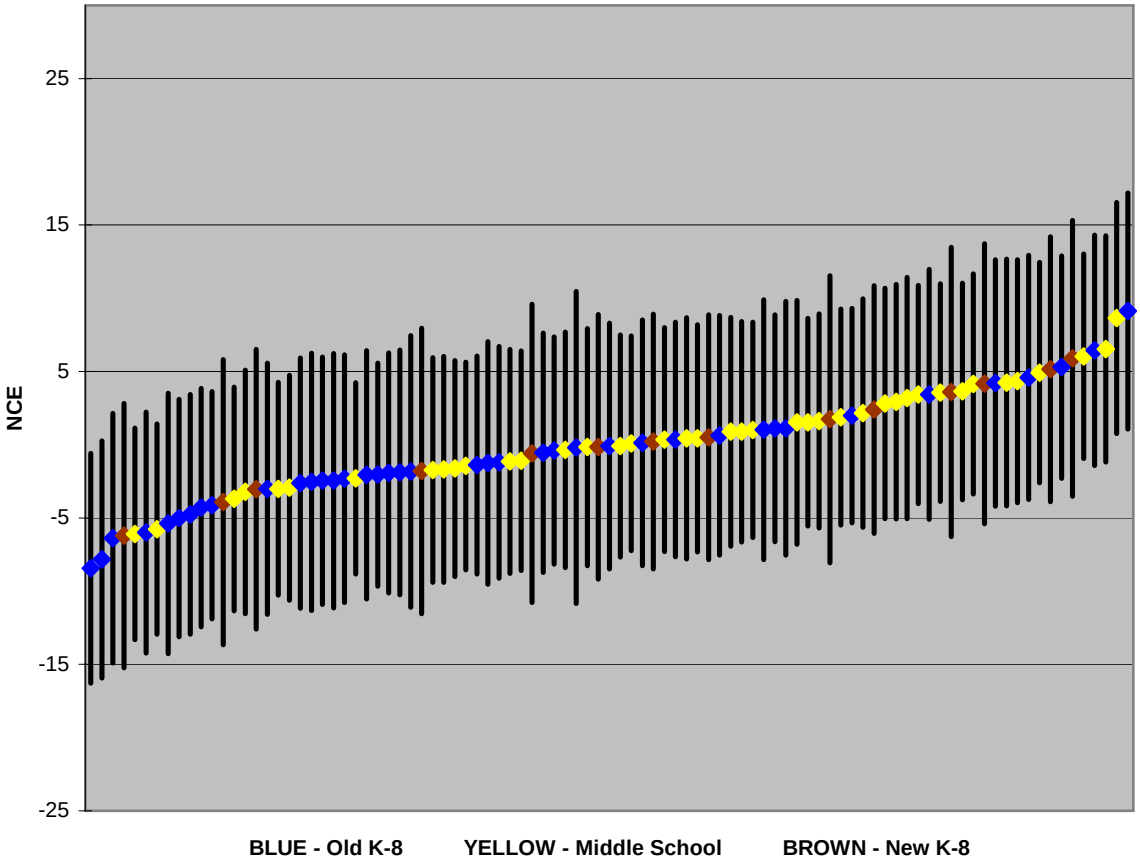


Chart 3
Level 3 Residuals, Full Model
Individual School Contributions to Mean Student Achievement
(with 95% Confidence Intervals)



References

Balfanz, R., Spiridakis, K. & Neild, R. (2002). *Will Converting High-Poverty Middle Schools to K-8 Schools Facilitate Achievement Gains?*. A Research Brief for the School District of Philadelphia. Philadelphia, PA: Philadelphia Education Fund.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years*. Boston: TIMSS Study Center.

Borman, G., Hewes, G. M., & Overman, L. T. (2003). "Comprehensive School Reforms and Achievement: A Meta-Analysis". *Review of Educational Research, vol. 73, no. 2, pp. 125-230*.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Newbury Park, CA: Sage Publications, Inc.

Burrill, G. (1998). "Changes in your classroom: From the past to the present to the future". *Mathematics Teaching in the Middle School, vol. 4, pp. 184-190*.

Coladarci, T., & Hancock, J. (2002). "The (Limited) Evidence Regarding Effects of Grade-Span Configurations on Academic Achievement: What Rural Educators Should Know". *Journal of Research in Rural Education, vol. 17, no. 3, pp. 189-192.*

Eccles, J. S. & Midgley, C. (1989). "Stage/Environment Fit: Developmentally Appropriate Classrooms for Early Adolescents". In R. E. Ames and C. Ames (Eds.) *Research on Motivation in Education, vol. 3.* New York: Academic Press.

Eccles, J. S., Lord, S. & Midgley, C. (1991). "What Are We Doing to Early Adolescents? The Impact of Educational Contexts on Early Adolescents". *American Journal of Education, vol. 99, no. 4, pp. 521-542*.

Epstein, J. L. & MacIver, D. J. (1990). *Education in the Middle Grades: Overview of National Practices and Trends*. National Middle School Association.

Herman, B. E. (2004). "The Revival of K-8 Schools". *Phi Delta Kappa Fastbacks, no. 519, pp. 7-37*.

Hough, D. L. (2005). "The Rise of the "Elemiddle' School". *School Administrator, vol. 62, no. 3, pp. 10-14*.

Jackson, A.W. & Davis, G. A. (2000). *Turning Points 2000: Educating adolescents in the 21st century*. New York, NY: Teachers College Press.

Kao, G. (1995). "Asian Americans as Model Minorities? A Look at their Academic Performance". *American Journal of Education, vol. 103, no. 2, pp. 121-159*.

Lee, V. E. & Smith, J. B. (1993). "Effects of School Restructuring on the Achievement and Engagement of Middle Grade Students". *Sociology of Education, vol. 66, no. 3, pp. 164-187*.

McEwin, C. K. & Dickinson, T. S. (1996). "Middle-level teacher preparation and licensure". In J. L. Irvin (Ed.), *What research says to the middle-level practitioner*. Columbus, OH: National Middle School Association.

McEwin, C. K., Dickinson, T. S. & Jenkins, D. M. (1996). *America's middle schools: Practices and progress – a 25-year perspective*. Columbus, OH: National Middle School Association.

McEwin, C. K., Dickinson, T. S. & Jacobson , M.G..(2005). "How Effectice are K-8 Schools For Young Adolescents?". *Middle School Journal, vol. 37, no. 1, pp. 24-28*.

Midgley, C. (1993). "Motivation and middle level schools". In M.L. Maehr & P.R. Pintrich (Eds.), *Motivation and Adolescent Development: Volume 8 of Advances in Motivation and Achievement,* (pp. 217-274). Greenwich, CT: JAI Press.

Mizell, H. (2005). "Grade Configurations for Educating Young Adolescents Are Still Crazy After All These Years". *Middle School Journal, vol. 37, no. 1, pp. 14-23*.

National Forum to Accelerate Middle Grades Reform. (2002, April). *Policy statement: teacher preparation, licensure, and recruitment*. Newton, MA: Author.

Offenberg, R. (2001). "The efficacy of Philadelphia's K-8 schools compared to middle grades schools". *Middle School Journal, vol. 32, no. 4, pp. 23-29*.

Paglin, C. & Fager, J. (1997). *Grade Configuration: Who Goes Where?*. Portland, OR: Northwest Regional Educational Laboratory.

Pardini, P. (2002)."Revival of the K-8 School". *School Administrator, vol. 59, no. 3, pp. 6-12*.

Peng, S. S. & Wright, D. (1994). "Explanation of Academic Achievement of Asian American Students". *Journal of Educational Research, vol. 87,July/August, pp. 346-352.*

Raudenbush, S. W. & Willms, J. D. (1995). "The Estimation of School Effects". *Journal of Educational and Behavioural Statistics, vol. 20, no. 4, pp. 307-335.*

Reising, B. (2002). "Middle School Models". *The Clearing House, vol. 76, no. 2, pp. 60-61*.

Schmidt, W. H., McKnight, C. C., Jakwerth, P. M., Cogan, L. S., Raizen, S. A., Houang, R. T., et al. (1999). *Facing the consequences: Using TIMSS for a closer look at the United States mathematics and science education*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Simmons, R. & Blyth, D. (1987). *Moving into adolescence: the impact of pubertal changes and school context*. New York, NY: Aldine De Gruyter.

Simmons, R., Black, A. & Zhou, Y. (1991). "African-American versus White Children and the Transition into Junior High School". *American Journal of Education, vol. 99, no. 4, pp. 521-542*.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel Analysis*. Thousand Oaks, CA: Sage Publications, Inc.

Weiss, C. & Kipnes, L. (2006). "Reexamining Middle School Effects: A Comparison of Middle Grades and K-8 Schools". *American Journal of Education, vol. 112, no. 2, pp. 239-272*.

Yakimowski, M. E. & Connolly, F. (2001). *An examination of K-5, 6-8, and K-8 grade configurations*. Report prepared for the board of School Commissioners. Baltimore, MD: Division of Research, Evaluation, & Accountability, Baltimore City Public School System.