Spring 5-17-2010

On Educational Outcomes in an Urban School District

Michael A. Gottfried
mgottfr2@wharton.upenn.edu

On Educational Outcomes in an Urban School District

Abstract
Urban school districts face an enormous challenge. They are confronted with high levels of poverty and minority students who at high-risk for educational failure. To compound this, financial resources are lacking in improving these dire conditions. Thus, in a situation where increased budgetary support is no longer accessible, one question remains: What will make a difference?

Chapter 1 suggests a first strategy. If district administrators or school principals could shift classroom composition to increase student achievement, then perhaps this managerial approach could improve urban education under extremely strict financial constraints. Using the framework of the education production function and two quasi-experiments, this investigation has identified status quo peer effects in Philadelphia's elementary school classrooms over six years of observations.

Holding fixed students and classrooms, Chapter 2 then asks what contributes to school effectiveness at the level of the institution. It does so by constructing two unique, quantifiable measures of school quality based on the empirical model from Chapter 1. The results indicate that institutional-level resources are significantly related to school quality across three categories (programs, personnel, and school environment) and within both testing subject areas.

Based on the covariates analyzed in the first two chapters, Chapter 3 evaluates if and why there is significant variation in standardized testing performance for students in a single urban school district. Incorporating variables into a three-tiered hierarchical linear model of student achievement explains the majority of the between classroom and between school variance, though only half of the within classroom variance.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Applied Economics

First Advisor
Robert P. Inman

Second Advisor
Rebecca Maynard

Third Advisor
Todd Sinai

Keywords
Economics of education, empirical economics, urban education, student achievement, peer effects, institutional effectiveness

Subject Categories
Education Policy | Other Economics | Policy Design, Analysis, and Evaluation | Public Policy | Urban Studies

ON EDUCATIONAL OUTCOMES IN AN URBAN SCHOOL DISTRICT

Michael A. Gottfried

A DISSERTATION

in

Business and Public Policy

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillments of the Requirements for the

Degree of Doctor of Philosophy

2010

Supervisor of Dissertation

*Signature*___________________

Robert P. Inman, Richard King Mellon Professor of Finance, Department of Finance

Graduate Group Chairperson

*Signature*___________________

Eric T. Bradlow, K.P. Chao Professor, Department of Marketing

Dissertation Committee

Robert P. Inman, Richard King Mellon Professor of Finance, Department of Finance

Rebecca Maynard, University Trustee Professor of Education and Social Policy,
Graduate School of Education

Todd Sinai, Associate Professor, Department of Real Estate

DEDICATION

For my family.

ACKNOWLEDGEMENT

I am privileged that Professor Robert Inman has been my adviser and dissertation chair. I am thankful for his many years of mentorship and support. His advice has extended beyond the scope of this dissertation. I also thank dissertation committee members Professor Rebecca Maynard and Professor Todd Sinai. Both have provided invaluable insight and guidance. The unique opportunity to work with these three economists in different schools and departments at the University of Pennsylvania has shaped my analytical framework and broadened by academic perspective.

I have had the opportunity to learn from other mentors during my tenure at the University of Pennsylvania, in both Wharton and the Graduate School of Education. They include: Laura M. Desimone, Fernando Ferreira, Douglas E. Lynch, Henry May, and Ruth Curran Neild.

I also thank current graduating students and graduated students from my doctoral program: Sam Chandan, Erica Johnson, and Walter Theseira.

Finally, I acknowledge my parents: Professor Adele Eskeles Gottfried and Professor Allen W. Gottfried. Both have taught me the meaning of persistence and dedication throughout my decades of education. As they are scholars in educational psychology and developmental psychology, I have immensely enjoyed and greatly learned from occasions when our disciplines have overlapped.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through GRANT R305C050041-05 to the University of Pennsylvania. The opinions expressed are those of the author and do not represent the views of the U.S. Department of Education.

ABSTRACT

ON EDUCATIONAL OUTCOMES IN AN URBAN SCHOOL DISTRICT

Michael A. Gottfried

Robert P. Inman

Urban school districts face an enormous challenge. They are confronted with high levels of poverty and minority students who at high-risk for educational failure. To compound this, financial resources are lacking in improving these dire conditions. Thus, in a situation where increased budgetary support is no longer accessible, one question remains: What will make a difference?

Chapter 1 suggests a first strategy. If district administrators or school principals could shift classroom composition to increase student achievement, then perhaps this managerial approach could improve urban education under extremely strict financial constraints. Using the framework of the education production function and two quasi-experiments, this investigation has identified status quo peer effects in Philadelphia's elementary school classrooms over six years of observations. This study further evaluated the potential growth in student learning from potential policies aimed at changing classroom composition. The results suggest statistically significant classroom peer effects on individual student reading and math achievement, though the effects differ based on a student's socioeconomic status.

Holding fixed students and classrooms, Chapter 2 then asks what contributes to school effectiveness at the level of the institution. It does so by constructing two unique, quantifiable measures of school quality based on the empirical model from Chapter 1. The results indicate that institutional-level resources are significantly related to school quality across three categories (programs, personnel, and school environment) and within both testing subject areas. While there is some consistency across school quality in reading and math, the results also indicate that differentiating between subject tests is crucial: school resources may provide distinctive institutional effects depending on the testing area itself.

Based on the covariates analyzed in the first two chapters, Chapter 3 evaluates if and why there is significant variation in standardized testing performance for students in a single urban school district. The initial results indicate that the overwhelmingly most significant contributor to total variance in achievement is within classrooms at the student level. However, incorporating variables into a three-tiered hierarchical linear model of student achievement explains the majority of the between classroom and between school variance, though only half of the within classroom variance.

TABLE OF CONTENTS

	Page
List of Tables	vii
List of Illustrations	x
Chapter 1: Peer Effects in Urban Classrooms: Evidence from Philadelphia	1
Chapter 2: Programs, People, and Property: Examining the Institutional-Level Factors of Urban School Effectiveness	90
Chapter 3: Assessing Access: Is there Educational Equity within Urban Schooling	134
Appendices	166
Bibliography	177

LIST OF TABLES

Page

Chapter 1
Table 1: Demographics of Urban Districts in the United States: 2008 – 2009 School Year 69
Table 2: Summary of Peer Effects Studies 70
Table 3: Student Panel Data Coverage by Cohort, Grade, and Year 71
Table 4: Descriptive Statistics of Full Dataset 72
Table 5: Qualitative Evidence from Former Principals, Principals, and Teachers 73
Table 6: Percentages of Grades per Year Demonstrating Evidence of Unequal Student Distributions 74
Table 7: Summary Statistics of Trimmed Samples 75
Table 8: Characteristics of Late Arrivals and Their Non-Late Classmates 76
Table 9a: Logistic Regression Results Predicting Probability of Classroom Assignment Based on Observable “Down-Branch” Characteristics in Illustration 1 77
Table 9b: Logistic Regression Results Predicting Probability of Classroom Assignment Based on Observable “Down-Branch” Characteristics in Illustration 1, Including Previous Test Scores 78
Table 10: Baseline Model of Education Production 79
Table 11: Peer Effects for the Baseline and Trimmed Samples 80
Table 12a: Estimating the Effect of Late Arrival Students on Non-Late Arrivals: Reading Achievement 81
Table 12b: Estimating the Effect of Late Arrival Students on Non-Late Arrivals: Math Achievement 82
Table 13: Peer Effects for Trimmed Sample, Broken Out by Free Lunch and Non Free Lunch Students 83
Table 14: Same-Year Impact on Learning by a One Standard Deviation Increase in a Classroom Trait (Within-District) 84
Table 15: Initial Impact: The Effect on Learning Months over Three Years (Within-District) 85
Table 16: Continual Impact: The Effect on Learning Months over Three Years (Within-District) 86
Table 17: Same-Year Impact on Learning of a One Standard Deviation Increase in a Classroom Trait (Within-School) 87
Table 18: Initial and Continual Impacts’ Effects on Learning Months (Within-School) 88
Table 19: The Effects of Peers on Non-Academic Outcomes 89

Chapter 2
Table 1: Descriptive Statistics of School-Level Production Function Inputs and Outputs 129
Table 2: Parameter Estimates of Logistic Regressions 130
Table 3: Regression Coefficients Predicting Reading School Effectiveness 131
Table 4: Regression Coefficients Predicting Math School Effectiveness 132
Table 5: Regression Coefficients and Effect Sizes Predicting Average School Behavior and Average School Absence Rate 133

Chapter 3
Table 1: Student, Classroom, and School Variables 161
Table 2: Partitions of Variance, Reading Achievement: Null Model 162
Table 3: Partitions of Variance, Math Achievement: Null Model 163
Table 4: Partitions of Variance, Reading Achievement: Full Model 164
Table 5: Partitions of Variance, Math Achievement: Full Model 165

LIST OF ILLUSTRATIONS

Chapter 1
Illustration 1: Resource-Based Assignment Process of Late Student (if assigned non-randomly) 68

Chapter 2
Illustration 1: The Total Effect of Assistant Principals on Reading School Effectiveness 128

Chapter 3:
Illustration 1: Reading Achievement for Schools in the Philadelphia School District Sample, 1995-96 through 1998-99 159
Illustration 2: Math Achievement for Schools in the Philadelphia School District Sample, 1995-96 through 1998-99 160

CHAPTER 1

PEER EFFECTS IN URBAN CLASSROOMS: EVIDENCE FROM PHILADELPHIA

Chapter Abstract

By first developing a theoretical understanding of student achievement under the education production function framework, this paper then empirically evaluates the effect of classroom peers on standardized test achievement for all elementary school students in the Philadelphia School District, over the school-years 1994/1995 through 2000/2001. With this unique individual- and multi-level dataset, two quasi-experimental strategies are employed. The first relies on the even distribution of students among classrooms within a particular grade and school in a given year. As a test of robustness, a second quasi-experimental strategy is employed, which depends on the idiosyncratic variation in classroom composition based on the random assignment of students entering the school at abnormal times during the academic year. Two sub-samples, differentiated by socioeconomic status, are subsequently evaluated to draw distinctions among the effects of classroom composition. Based on these strategies, this study finds statistically significant classroom peer effects on standardized achievement and additional months of learning, though the degree to which they impact performance differs based on socioeconomic status.

Introduction

Urban school districts are in a bind. They are faced with high levels of poverty and minority students who are underprepared for postsecondary opportunities or are at high-risk for educational failure (Tighe, Wang, & Foley, 2002). To compound this, financial resources are lacking in improving these dire conditions (Lewis, Baker, Jepson, et al., 2000). Thus, in a situation where increased budgetary support may not be accessible, a question remains: Is it possible to improve educational outcomes for urban youth without spending an additional dollar?

This paper suggests one such strategy. If district administrators or school principals could shift classroom composition to increase student achievement, then perhaps this managerial strategy could improve urban education under extremely strict financial constraints. This study assesses status quo peer effects in urban classrooms and the potential growth in student learning from policies aimed at changing classroom composition. While this paper examines these effects within the context of Philadelphia, the issues are relevant to many urban districts handling similar demographic populations, as seen in Table 1.

Peer effects are important in schools because they provide insight as to whether students can be affected by the achievement and other characteristics of their classmates. In fact, parents, educators, and researchers have long believed that peer quality is one of the most important determinants of student outcomes (Henderson, Mieszkowski, & Sauvageau, 1977; Link & Mulligan, 1991; Summers & Wolfe, 1978; Zimmer & Toma, 2000). However, few empirical studies have successfully captured the impact of peer groups on student performance. Further, the evidence that does exist is often conflicting or has been open to varying interpretations. This paper employs two unique identification strategies – the first based on a sample of evenly distributed classrooms within a grade and the second based on exogenous disruptions to classroom composition – in order to develop unbiased estimates of the effect of peers on student achievement.

Since 1966, when Coleman and his authors (Coleman et al., 1966) provided a first in-depth perspective into the relationship between classroom composition and subsequent achievement, a vast literature in sociology, psychology, economics, and education policy has burgeoned around the question of whether better peers can lead to even better outcomes. From a sociological and psychological perspective, the peer group can be an important source of information and motivation. Sociologically, students influence each other by learning in groups, potentially helping one another and discussing classroom concepts and perspectives. Because a peer group can enable discourse among students, this sociological unit provides a mechanism for processing new information and hence disseminating different interpretations, thereby advancing cognitive ability. Psychologically, peers act as important role models, which are seen as powerful means of transmitting attitudes, values and patterns of thought and behavior (Bandura, 1986).

The effects of peer groups are as pertinent to researchers in economics and education policy. First, social interactions among students can be interpreted as creating positive and negative externalities. That is, peer groups can induce spillover effects in classroom learning through productive or disruptive behavior (Lazear, 2001). As such, it is crucial to uncover and further understand the significance of peers and their implications on classroom productivity. Doing so will enable policy makers to determine which inputs matter in educational school reform, thereby providing insight on how school inputs make a difference in the classroom rather than on just whether or not they do at all. Second, a major question in the economics and policy literature is whether or not the interactions among students lead to large social multipliers (Epple & Romano, 1998; Hoxby, 2000). Depending on the nature of peer effects, there may be gains from grouping together different subsets of students. Answers to these questions would inform the debates on school choice, busing, and tracking (Angrist & Lang, 2004).

In this paper, two conditions are accepted as given. First, parents sort according to Tiebout (1956) and thus choose a school based on neighborhood and peer characteristics. Second, schools can manipulate the assignment of students into classrooms. Nonetheless, two empirical strategies are implemented in which, even under these two conditions, allow for estimates of peer effects that have minimized selection bias.

The first strategy relies on the distribution of the observable characteristics of students among classrooms within a grade and school for a given year. It is possible that a school administrator may sort students by classrooms in a given grade, thereby creating rooms of unequal weight (and thus implementing a tracking policy). However, this study supports the notion that there are also classrooms in many grades (particularly within elementary schools) which remain more heterogeneously distributed (an artifact of potential random assignment of students). This paper will thus examine a sample in which student traits in particular classrooms within a given grade are equally distributed. This will be labeled as the "trimmed" sample. As such, this strategy provides a more refined basis for evaluating peers and hence evades the estimation errors associated with examining classrooms with uneven distributions (i.e., tracked) of student traits. From this evenly trimmed sample, two sub-samples of are evaluated. The first is comprised of non-special needs, students receiving free lunch, and the second of non-special needs non free lunch recipients. These samples, differing only in terms of socioeconomic status (SES), can provide insight into how peer effects impact two distinct socioeconomic groups of elementary school students within an urban school district.

To test the robustness of this first method, a second strategy is employed. This second approach depends on the idea that there is some degree of idiosyncratic variation in classroom composition, based on the assignment of students entering the school year at abnormal times at least one month after the beginning of the school year. This strategy will be referred to as the "late arrivals" strategy. Even though schools (or parents) may make active decisions regarding classroom placement, this strategy upholds that the assignment of a late arrival is random and hence affects the classroom peer group above and beyond the school’s control. Furthermore, because this late arrival is the final student to be assigned to the classroom, this second method can shed light on marginal peer effects.

Based on these two empirical strategies, this paper finds statistically significant classroom peer effects on individual student reading and math achievement. The trimmed sample demonstrates that peer effects are evident in classroom academic ability, both in reading and math. Classroom peer effects are also prevalent in other channels of peer characteristics, including student behavior and gender. The results differ when the samples are broken out by free lunch recipient status, demonstrating that free lunch students are slightly more at risk by the negative aspects of peer effects. The late arrivals strategy confirms the results from the trimmed sample. Because of the similarities in sizes and magnitudes of the coefficients between the two strategies, this approach suggests that the marginal effect is similar to the average peer effect. Hence, not only are the peer effects robust throughout the analysis, but the consistency in results also allows for a policy discussion to be had around changing classroom composition.

Peer Effects Literature

Under the rubric of the economics of education, a common thread within the literature on peer effects is the production function. The education production function models the relationship between school inputs and various measures of achievement as if learning were comparable to a firm’s production process. Education production studies attempt to determine the relationship between inputs measured by neighborhood, family, teacher, school, and classroom characteristics with output measured by student achievement. Relevant education production function literature to this paper will have evaluated the relationship between classroom composition and student-level achievement.

Education production studies relating to peer effects surfaced in 1966 with Coleman's Equality of Educational Opportunity report (Coleman et al., 1966). In this seminal piece, the authors determined that family and regional backgrounds were the overwhelming factors of student achievement. To evaluate peer effects, this report focused on the composition of black and white students in the classroom as primary correlates that could affect a student's achievement. From this analysis, the authors reported higher achievement levels for disadvantaged black students who attended middle class schools, thereby allowing this research to attribute race (e.g., student background) as the significant factor in achievement. However, this study did not control for self-selection: students and families chose to be part of the experimental subset of economically disadvantaged black students attending wealthier schools. That is, those black students who found themselves at higher quality schools did not end up there randomly, and perhaps the results of this study were confounded.

Along the same lines, Hanushek (1972) also utilized student and peer racial backgrounds as production function input measures of current achievement outcomes. In this study, the author attempted to determine a relationship between the varying proportions of black students in a classroom and the subsequent effect on student achievement. Like the *Coleman Report* (1966), this early paper focused on racial classroom composition as the primary determinant of peer effects. In essence, the peer effect was not defined as embodying other components, such as interacting with students of different ability, but was rather solely reliant on race.

Although the importance of researching peer effects developed within the literature in 1960's and early 1970's, little quantitative work had been conducted up to that point. However, in the late 1970's two major empirical studies on the determinants of peer effects were disseminated into the field (Henderson, Miezkowski, & Sauvageau, 1978; Summers & Wolfe, 1977). Using more rigorous empirical techniques, these education economists began to incorporate other aspects of education production into the determinants of peer effects, in addition to those family (i.e., race) and socioeconomic factors previously studied.

Although these two papers presented different findings and mixed results, they are still nonetheless considered to be significant in the field for having been first to apply quantitative rigor to the evaluation of classroom peer effects. Summers and Wolfe (1977) found that both high and low ability students benefit from an academic improvement in the peer group. However, the effect was largest for low ability students. Henderson, Miezkowski, and Sauvageau (1978) also found that students of all abilities were affected by their peers, but that the benefit from an improvement in the peer group was independent of student ability. In addition, peer effects were possibly nonlinear, implying that student performance rises with average classroom IQ, but that this increase slows as the average itself increases.

Jumping ahead several decades, most contemporary studies on classroom peer effects have had the luxury of the availability of larger and increasingly more detailed datasets; this is a consequence of improved schools' record keeping and reporting requirements. Nonetheless, even with an expanded set of empirical resources, the results in the literature are mixed. There still remains very little consensus over the effects of classroom composition on student achievement.

This lack of consensus seems to stem from the fact that uncovering unbiased estimates of peer effects is not an easy task. The reason is that there are two major confounding issues present among empirical studies on classroom peer effects: first is the self-selection by families into neighborhoods and schools, and second is the non-random assignment of students into classrooms by school management. In more detail, first, families self-select into schools based on various characteristics including income and residential and educational preferences. As such, families do not randomly assign themselves to neighborhoods but rather intentionally do so according to tastes and resources (Tiebout, 1956). As a result, school and family backgrounds are confounded with classroom characteristics and hence with peer effects. To provide evidence of this, Jencks and Mayer (1990) showed that the magnitude of estimated peer effects tended to decline as more controls for parental characteristics were included. Of course, it would be ideal to have family information, as Jencks and Mayer (1990) demonstrated. However, family attributes are often omitted from analyses, simply because they are not contained within the administrative data on student and school characteristics. To overcome this problem, this current paper will utilize lagged-test scores in a value added model for each individual student per year, which serves as a proxy for an individual fixed effect. This will mitigate confounding issues relating to current achievement and unobserved family background characteristics.

Driving a second concern in estimating peer effects, school management creates selection issues. That is, schools may assign students to particular classrooms based on specific observable student attributes, such as a previous year's behavior grade. A school principal may hypothetically assign all poorly behaved students in a grade to a particular classroom. As such, within these classroom, there are unobserved factors affecting the contemporaneous achievement of both individual students and their peer groups. This selection issue is what Manski (1993) named as the reflection problem, in which it is difficult to distinguish between the effects of individual-level student factors and those from the peer group. Empirically, if any of these student characteristics have positive or negative effects on achievement, then the estimates will be biased. This selection bias will be overcome in this paper by using lagged measures of peer group achievement as well as a tracking assignment algorithm, removing the sample of tracked classrooms and hence potential non-random peer group formation.

Often, the research ignores these two issues, as Moffitt (2001) has noted. However, some noteworthy papers have attempted to overcome these confounding statistical issues with quasi-experimental methods. But even with these more finely-tuned quantitative contributions, the literature still remains inconclusive on the effects of peers. Table 2 presents a summary of these recent studies on peers.

Evans, Oates, and Schwab (1992) attempted to identify instrumental variables that are correlated with peer effects but not with unobserved determinants of achievement, such as family sorting or classroom placement. Their results were mixed, depending on the empirical method employed. Using differences in school quality induced by residential location and magnet school lotteries, Cullen, Jacob, and Levitt (2003) found no academic benefit associated with attending a school with better peers. Hoxby (2000) utilized the exogenous changes in demographic and gender composition of contiguous elementary school grade cohorts to evaluate the effect of peers. She determined that peer effects do play an important role on achievement, particularly within gender differentials. Angrist and Lang (2004) found that exogenous changes in classroom composition have at most transitory effects on the achievement of minority students.

Hanushek, Kain, Markman, and Rivkin (2003) attempted to overcome the confounding issues of omitted family variables through the use of a fixed effects framework and lagged measures of peer achievement. The authors reported positive influences of higher achieving peers. McEwan (2003) also utilized an identification strategy based on fixed effects and found that the mean schooling of mothers in a classroom provides a strong link to class achievement. However, Ammermüller and Pischke (2006) found evidence of peer effects, as in the two former studies, but that they drop away once non-linear dimension was taken into account. All three of these papers utilized school fixed effects and compared students in different classes to help circumvent the self-sorting problem of students and schools. However, this method on its own requires an assumption that students not be sorted in different classrooms according to their ability levels. This would violate the second confounding issue mentioned above.

To avoid management selection bias, Figlio (2005) introduced a unique identification strategy to estimate peer effects. He used the fraction of boys with female-sounding names in a classroom as an instrument for peer behavior. Figlio's study found that peer disruptive behavior was associated with an increased likelihood that other students were suspended in the class and a decreased likelihood of improved academic achievement.

In the realm of higher education, peer effects are often studied using natural experiments. For instance, Sacerdote (2001) and Zimmerman (2003) estimated the peer effects of randomly assigned college roommates at Dartmouth College and Williams College, respectively. Sacerdote (2001) found that roommate peers had an impact on grade point average and decisions to join social groups. Zimmerman (2003) found that a student with a low or middle-range SAT score negatively affected a roommate with an also low SAT score. That being said, the academic atmosphere of college roommates differs from elementary peer effects. The generalizability to other realms of education policy remains potentially dubious.

This paper differs from the peer effects literature in three capacities. First, this study puts forth two quasi-experiments in the elementary school context in which it is possible to track students at the individual, classroom, school, and neighborhood levels. Many studies in the literature use aggregate data (e.g., grade peers rather than classroom peers) or attempt to deal with selection issues by simply controlling for observable characteristics. However, these studies remain unconvincing because observable characteristics remain correlated with unobserved selection and assignment (Rothstein, 2008). As such, quasi-experimental methods are necessary to provide for the random assignment of students. In addition, many studies focus on middle or high schools. However, only in elementary school do students spend most or all of their time in a single classroom and hence with a single peer group. Once students enter middle and high school, they move around throughout the day and are susceptible to the influences of many peers (Betts & Zau, 2004).

Second, having individual- and multi-level data, this study can draw distinctions among multiple channels of peer effects, based many on observable characteristics. Moreover, it is possible differentiate between the classroom peer effects on students of differing socioeconomic status, an aspect that the literature has not yet assessed.

Third, the second empirical strategy allows for the evaluation of the effects of the random assignment of single late (and therefore last) student placed into a classroom. Hence, this quasi-experimental method enables for the estimation of marginal peer effects based on a multitude of student characteristics. The determined linear relationship of changing peer composition allows for a policy discussion to follow.

The Education Production Function

To examine peer effects, this study employs the education production function, as initially developed by Summers and Wolfe (1977) and Henderson, Mieszkowski, and Sauvageau (1978), and later revised by Todd and Wolpin (2003). This model has consistently served as the foundation for evaluating the effect of peers on academic achievement. The "output" is standardized test score performance, as determined by a set of "input" vectors consisting of a wide range of independent variables.

Rather than assuming that a current year's achievement outcome is strictly a function of current inputs, it is possible to enrich the education production function model to include inputs from previous time periods. In fact, it is theoretically possible to include all time periods for which the student is in school. This model is known as the historical model of education production. To derive this full historical, cumulative-learning model, it is important to make an initial assumption, as developed by Todd and Wolpin (2003): achievement in the initial period of schooling is a function of the student's natural endowment and family inputs provided prior to the period in which the student enters his or her first year of schooling. Those family inputs in the previous period of initial schooling are described as follows:

$$F_0 = f_0(G_i), \tag{1}$$

where F_0 is family inputs in before-schooling period 0 and G_i is student *i*'s natural endowment. Because the student has not yet enrolled in school in period 0, there is no academic achievement information for the student. Hence, the family at this point can only adjust its inputs to the student's learning process based on their direct observations of the student's ability level, G_i.

Then, in the first period of schooling, student achievement is a function of ability G, family inputs F, and contemporaneous school inputs S:

$$A_1 = f_1(G_i, F_0(G_i), F_1(G_i), S_1) \tag{2}$$

Note that in this first year of school learning, school inputs do not adjust to the child's ability. In practice, this is demonstrated by the fact that students are more-often-than-not randomly assigned to a classroom in the starting grade that the school offers, either kindergarten or first grade.

In subsequent periods, however, schools and parents can potentially adjust their respective inputs, based on the student's reported achievement performance from the previous period. This may be realized in some schools as tracking, and is demonstrated by the education production function in year 2 of schooling. Family and school inputs in year 2, respectively F_2 and S_2, are functions of A_1, the previous year's achievement:

$$A_2 = f_2(G_i, F_0(G_i), F_1(G_i), F_2(G_i, A_1), S_1, S_2(A_1)) \tag{3}$$

Iterating this process for each year of schooling provides the following education production function for a student in a given year of school *t*, which includes both contemporaneous and historical information:

$$A_t = f_t(G_i, F_0(G_i), F_1(.) \ldots F_t(.), S_1, S_2(.) \ldots S_t(.)) \tag{4}$$

This model states that achievement, for a student in a given year *t*, is a function of a student's natural endowment (which does not change over time), the family's inputs in the year prior to schooling and through year *t*, as well as school inputs from the first year of schooling through year *t*.

With equation (4), it is possible to restate school inputs to include teacher and classroom components, which demonstrates the problem of biasing the achievement measure – school, teacher, and classroom inputs are all a function of previous achievement. In general, the literature assumes that it is possible to describe a linear relationship between the inputs and outputs of the education production function. A linear historical model built upon the concepts of equation (4) looks as follows:

$$a_{ijkt} = \beta_0 + \beta_1 G_i + \beta_2 F_{it} + \beta_3 N_{it} + \beta_4 S_{jkt} + \beta_5 T_{jtk} + \beta_6 C_{jt} + \beta_7 P_{-ijkt} + \sum_{t=1}^{t-1} \lambda_8 F_{it} + \sum_{t=1}^{t-1} \lambda_9 N_{it}$$
$$+ \sum_{t=1}^{t-1} \lambda_{10} S_{jkt} + \sum_{t=1}^{t-1} \lambda_{11} T_{jkt} + \sum_{t=1}^{t-1} \lambda_{12} C_{jt} + \sum_{t=1}^{t-1} \lambda_{13} P_{-ijkt} + \varepsilon_{ijtk} \quad (5)$$

where achievement a is for student *i* in classroom *j* in school *k* in year *t*; G is ability level for student *i*, which is unchanged by classroom *j* in school *k* in year *t*; F is a function of family inputs for student *i* in year *t*; N includes neighborhood characteristics for student *i* in year *t*; S are school characteristics, which the student experiences via classroom *j* in school *k* in year *t*; T are teacher effects in classroom *j* in school *k* in year *t*; C are classroom-specific characteristics for classroom *j* in year *t*; P are peer effects, derived from other students (i.e., "not *i*"); and the error term ε includes all unobserved determinants of achievement.

This linear representation in equation (5) separates current and historical inputs. However, it is a difficult and challenging task to acquire all inputs to estimate a fully historical education production function. One solution to this problem is to take the difference of equation (5) with respect to year *t*, the current year of schooling, and equation (5) with respect to year *t-1*, the previous year of schooling. The result is known as the value added specification, where all input requirements reduce to current inputs plus achievement from the *t-1* period:

$$\mathrm{a}_{\mathrm{ijkt}} = \beta_0 + \beta_1 F_{it} + \beta_2 N_{it} + \beta_3 T_{jtk} + \beta_4 C_{jt} + \beta_5 P_{-ijkt} + \beta_6 a_{ijk(t-1)} + \gamma_{\mathrm{ijtk}} \quad (6)$$

This model strictly incorporates current achievement, prior achievement, and contemporaneous inputs. Through the process of differencing the current year from the previous year, the value added model assumes that prior achievement captures the influences of all historical, noncurrent inputs. The model also assumes that learning in year *t* is reflected in year *t*'s achievement; in other words, there are no delays in the actualization of what is learned in a current school year. As a result, current achievement is not confounded with omitted characteristics that persist in prior periods of schooling (Hanushek et al., 2003). Note previous achievement is on the right-hand side of the equation. Unlike a model where the left-hand side variable is a difference between current and previous achievement, the approach utilized here does not constrain the parameter of achievement to be a value of one (Rothstein, 2008; Todd & Wolpin, 2003).

Because of the difficulty and even impossibility of quantifying the underlying, true measure of student ability (Hanushek, 1979), the value added model has a key feature of removing innate ability from the equation, as it is assumed here that unmeasured ability remains constant over time and is hence subtracted out via differencing the historical model of year *t-1* from year *t*. A further assumption of the model is that unmeasured student ability does not interact with any other covariates differently over time. It may be possible that the relationship between unmeasured student ability and peer characteristics, for instance, may change over time (i.e., higher ability students are boosted more by their peers as they progress through school); however, the current model has assumed the marginal effects of these interactions to be constant over time.

For the purposes of this evaluation, it is in the error term where school effects are identified. From equation (6), equation (7) decomposes the error term into four components:

$$\gamma_{ijtk} = \delta_k + \omega_t + \upsilon_{kt} + \varepsilon_{ijkt} \tag{7}$$

where δ_k are school fixed effects, ω_t are year fixed effects, and υ_{kt} are school-by-year fixed effects. Additionally, ε_{ijkt} is a random error capturing two additional pieces of information: a classroom-specific random component that is common to all members of that same classroom in a given year and individual shocks that vary over time.

School fixed effects account for sorting into school district catchment areas by comparing children from different classes within the same school. In essence, school fixed effects control for the average differences between schools. Similarly, year fixed effects control for average differences between years (e.g., an unseasonably cold winter drives down attendance in the district). To account for shocks during a given year, year fixed effects are thus also included. However, only do school-by-year fixed effects account for unobserved changes in the school environment as a student progresses through years of schooling (e.g., gentrification of an urban school's neighborhood, changes to school leadership, new curriculum, etc.). Thus, with this particular error structure, the empirical model in this paper allows for families and students who enroll in a particular school in a particular year to share similar preferences and unobserved characteristics, as they are e captured by the fixed effects components. Finally, idiosyncratic unobserved individual characteristics can vary across individuals.

Data

The analysis of classroom peer effects for this study is facilitated by a unique and comprehensive dataset of student, teacher, and neighborhood observations.[1] Student and teacher data were obtained from the School District of Philadelphia via the District's Office of Student Records and through the District's Personnel Office. Neighborhood data were obtained from the 2000 Census flat files at the census block level. Neighborhood data relating to age, sex, households, families, and housing units were merged from the Census Summary File 1; additional social, economic, and housing measures were merged from Summary File 3. The data sample in Summary File 3 includes one in six households that received the long-form Census survey, whereas Summary File 1measures are based on the full universe of responding households.

[1] A full description of variables used in this study can be found in Appendix A.

Student Data

The student data is organized into five panels, or cohorts, of students. The grade-year progressions for each cohort are described in Table 3. The first three cohorts (A, B, and C) have observations in the dataset starting in the 1994/1995 academic year, while cohorts D and E are the kindergarten classes in the 1995/1996 and 1996/1997 academic years, respectively. Each cohort consists of approximately 16,000 students in each year with the exact number changing from year to year as students enter and leave the school system (or change cohorts due to grade retention or advancement).

Student can be tracked throughout their tenures in the Philadelphia School District. Students cannot be tracked if they leave the school system – no information is available for students who leave the District for other districts, private, or parochial schools. However, because students retain their unique identification numbers in the District's record system, if students should return into the District, they can be matched back to their original records. Because of this intricate tracking mechanism of incoming, outgoing, and returning students, the sample includes the entire population of cohorts of students in the Philadelphia public school system.

The shaded area of Table 3 describes the data to be used in this study. Grades 5 and higher are truncated from the sample because the dependent variable is a measure of a standardized student achievement score, for which only grades 2 through 4 are available.

Nonetheless, the five cohorts of elementary school students are represented in the analytical sample.

For each student in each academic year, basic information concerning personal characteristics such as date of birth, gender and race is augmented by a rich selection of variables in three categories. First, performance variables include: teacher-assigned behavior grades;[2] and Normalized Curve Equivalent (NCE) scores in math and reading from the Stanford Achievement Test Ninth Edition (SAT9) for grades 2 through 4.[3] Second, students are identified as: special education; English language learning (ELL) student; free lunch recipient; and having been enrolled in kindergarten within the Philadelphia School District. Third, school, grade, and room assignments are available for each student per year.

In addition, information was collected on a student's home address, including street number and name and zip code. The merging of neighborhood data was achieved by geo-coding each address to its longitude and latitude and then assigning each student to a census block group. Just under 94-percent of the students were successfully geo-coded and mapped to their respective block groups. Without family information, the vector of neighborhood variables serves as proxies for unobserved family characteristics in

---

[2] Behavior is assigned as a letter or number grade on a student's official record at the end of the academic year. The rubric is based upon a student's ability to demonstrate responsibility, get along and show respect for others, respect materials and supplies, follow rules, and show appropriate citizenship in the classroom and in other areas.

[3] The NCEs are the generally preferred measurement for methodological reasons – they have statistical properties that allow for evaluating achievement over time (Balfanz & Byrnes, 2006). Normal curve equivalents range in value from 1 to 99 with a mean of 50.

empirical models (e.g., Hanushek et al., 2003). Free lunch is the only indicator based on direct observation of family characteristics (e.g., total household income). The free lunch indicator implies that a student's family income is less than 130-percent of the federal poverty guideline, accounting for family size. More than half of the students in the Philadelphia School District are classified as free lunch eligible.

Table 4 describes the student data for two relevant populations. The first column describes the entire student population for all cohorts over all academic years, based on Table 3. The population, narrows, however, when the data is confined to the relevant grades and test scores in a value added specification (i.e., difference in which second grade data can only be used as a lagged test score for a third grade student), as seen in column 2. In addition to the requirement of having all test score information, the data in column 2 are restricted by missing student information, lacking teacher data, and class size restrictions. As consistent with Ammemuller and Pischke's (2006) data truncation methodology, any classroom in this dataset that has fewer than 12 students was removed from the analytical sample. Note that there have been multiple iterations of a random sample drawing of students from both larger and smaller samples in order to conduct a test of mean differences. The t-statistics, based on this random sampling algorithm, are not significant – there are no structural differences between the full population and column 2, the analytical sample.

Teacher Data

Data on teachers comes from both from student records and from the District's Personnel Office. The student record provides the name of the teacher assigned to a student's classroom in the given academic year. In addition, a more detailed dataset on teacher characteristics was obtained from the District's Personnel Office.

From these sources, four sets of variables were incorporated into the dataset. First, for each teacher, basic characteristics include race and gender. Second, a measure of teacher experience is based upon appointment date variables, including district appointment date, teaching seniority date, and present position appointment date. Third, a binary variable indicates whether a teacher had a Master's degree, based on the record which provides detail on which graduate school the teacher had attended. Finally, a binary variable indicates if a teacher had received Pennsylvania state certification, based on completion of either Level I or Level II Certificates.

The Student-Teacher-Classroom Observation

Table 4 also presents corresponding teacher and classroom data for each student in the database. The variables presented in this table are based on the teacher data files and a student-teacher-classroom matching algorithm. Students can be grouped unambiguously into classrooms because of the school and classroom assignment information in the student database. In contrast, the teacher dataset does not include school or classroom assignment. Teachers are matched to their classrooms by matching their name, as it appears on their personnel record, to the teachers' names as it appears in the dataset. The name of each student's teacher in each year appears as part of the student's record; that information is extracted from the student's report card along with the classroom number. The name that appears on the report card is not always the full name of the teacher, and thus the matching algorithm is conservative in requiring that teacher surname and given name must both be matched to be considered a correct student-classroom-teacher observation.

Because each student observation includes the school, grade, and classroom assignment of the student in each academic year, there is sufficient information to assemble classroom data. The peer characteristics for each classroom include summary statistics of the characteristics of the students in the classroom.

Identification Strategies

Within the literature, the primary obstacle in identifying peer effects has been that students are not randomly allocated to either schools or classes (Rothstein, 2008). When students are intentionally assigned to rooms, a student's peer group likely correlates to his or her own unobserved ability and motivation, which in turn correlates to his or her testing performance. As a result, the estimate of the peer effect is biased due to these joint unmeasured correlations. In order to have a sufficient identification strategy of peer effects then, variation in peer composition that affects classroom outcomes must not be correlated with those unobserved determinants of classroom outcomes (Manksi, 1993). If students are indeed randomly allocated to classrooms, then peer effects will not be correlated with unobserved determinants of student ability. Thus, if there is any relationship between peer group, unmeasured ability or motivation, and achievement, then it has only occurred randomly.

The identification strategies described in this paper exhibit this feature by implementing two quasi-experimental approaches based on the random assignment of students to classrooms. Using the random assignment of students to classrooms will break the link between peer characteristics and unobserved influences on the classroom. In conjunction with these quasi-experimental methods, using a value added model of student achievement has also reduced the correlation between student outcomes and omitted measures (i.e., ability). However, the possibility still exists that determinants of student outcomes remain correlated to unmeasured student ability even after employing the quasi-experimental methods with a value added model specification. Other empirical methods are hypothetically possible, such as the use of instrumental variables which would reduce this bias in the peer effects estimates. However, without an appropriate instrument of unmeasured ability, the implementation of the value added model on a randomly assigned set of students remains the most robust methodology.

This study has employed the two quasi-experimental strategies on third and fourth grade classrooms within the Philadelphia School District. Though at the cost of losing some degrees of freedom, restricting the analysis to these particular grades allows for the strict use of standardized test scores as an educational outcome in a value added empirical model. Moreover, this study strictly relies on elementary school classrooms because students remain in the same room through the school day. Once students begin middle school, classes (and hence peer groups) alternate so much throughout the day as the student goes from period to period that peer groups from one class potentially become too muddled (Betts & Zau, 2004). Furthermore, limiting the sample strictly to elementary school students avoids the selection bias issues relating to drop outs in high school years (Ehrenberg & Brewer, 1994).

Strategy 1: The Trimmed Sample

The first strategy depends on the observed distribution of students in classrooms in a grade-school-year unit. That is, as long as students are evenly distributed among classrooms, then the peer effects estimates will not be biased by unequal classroom compositions that often occur under tracking policies (Ammermueller & Pischke, 2006; Manksi, 1993; Rothstein, 2008).[4] Equally distributed students in a given grade in a school-year will be deemed part of the "trimmed" sample. This will provide a measure of the average effect of peers in the average classroom within the district. If, on the other hand, there are statistically significant differences in observed characteristics of students within the classrooms of a particular grade, then all students in that grade will be

[4] A typical example would be an average ability student in a classroom of predominantly high ability students. This may increase the average student's motivation (and hence testing performance) simply because the student is aware that he or she has been placed in a higher performing classroom.

considered tracked for the purpose of this analysis.[5] They will not be included into the analytical sample. The results from using observations from these unevenly distributed tracked classrooms would be confounded with other factors, such as management's manipulation of the classroom peer environment in a way that non-randomly correlates a student's peer group to unobserved ability and subsequent testing performance. As such, the estimates based on these tracked grades would be statistically biased, and thus all classrooms within the unevenly distributed grades-per-school must be removed from the sample.[6]

Hence, what is necessary for a more unbiased evaluation, then, is the identification of a population of students who are in evenly assigned classrooms in grades in order to emulate random assignment. Monk (1987) reports that elementary school principals often randomly assignment students from one year to the next. In an interview with one elementary school principal regarding student assignment, Monk noted the following in his paper: "As the principal put it: '(It's) just very random, no real look at any criteria for the simple reason that sometimes at the elementary levels that's the best kind of grouping" (p. 170). Furthermore, Monk (1987) reported that socioeconomic status and principal involvement in the assignment process of students were directly related. A

[5] It is possible that a principal may track students so that each room has an even distribution of students (based on a given trait). However, this policy would not bias our estimates because students would not be assigned to a room that is observably different from the others in a grade from the student's perspective (i.e., if there are 3 behavior problems per room, then a non-behavior problem would not feel as though he or she was assigned to a behavior problem room).

[6] Though it is possible that statistically significant unequal distributions of student characteristics occurred by chance and removing them does reduce the sample more than necessary, doing so nonetheless ensures that extreme classroom compositions will not bias the estimates of peers.

lower SES at an elementary school implied less principal involvement. Ammermuller and Pischke (2006) also found lack of intentional tracking processes within elementary schools. Thus, there appears to be evidence within the literature that elementary school students, particularly in low SES urban school districts like Philadelphia, are randomly assigned.

There also seems to be evidence among practitioners. This study conducted a series of interviews during the 2009/2010 school year of principals, former principals, and teachers in the School District of Philadelphia. The results of these interviews suggest that students are heterogeneously assigned to classrooms, most often by simply assigning rooms to be 50-percent of each gender. Moreover, students of a particular characteristic or trait (e.g., behavior problem) are not assigned homogeneously to a single room in a given grade. Instead, the evidence suggests that types of students are distributed evenly across classrooms in that grade. What results, then, is an even distribution of student characteristics across rooms. Table 5 presents evidence from these interviews, suggesting that intentional tracking policies generally appear to be absent in this District.

If and when students are placed homogeneously (i.e., tracked) within a grade, however, the literature on assignment consistently agrees that academic ability and socioeconomic status play a major role in this intentional classroom placement. Argys, Rees, and Brewer (1996) found that after holding socioeconomic status constant, race and ethnicity were not significant indicators of tracking placement. They hypothesized that race, as an observable characteristic, may be confounded with other determinants of placement.

This study has evaluated these observable characteristics in the sample of elementary school students in Philadelphia in order to precisely identify a sample of evenly distributed classrooms within grades based on academic ability, socioeconomic status, and gender. Specifically, these characteristics include: previous year's academic performance; behavior grade (i.e., a previous year's "D" report-card grade determines a student to be a behavior problem); free lunch, English language learner, and special education status; and student gender.[7]

Because of the individual- and multi-levels of the data used in this analysis, this study can identify both student characteristics and the overall observed classroom characteristics in each academic year. As a result, it is possible to evaluate the distribution of observable characteristics for classrooms in a grade-school-year unit of observation. Determining which grades in a school-year exhibit evidence of unequal weights (i.e., what will be deemed as potential tracking) requires the use of one-way analysis of variance (ANOVA). One-way ANOVA is used to test the differences among two or more groups, testing the null hypothesis that the population means are equal. The one-way ANOVA test is implemented and iterated for each grade, per year per school. When the F-statistic

[7] An additional characteristic – whether the student was repeating the current year's grade – was included in a version of the model. However, the classroom peer variables pertaining to retained students were insignificant (though negative). All other peer characteristics mentioned above remained robust to the inclusion of repeaters in the model. Therefore, the peer effects analysis of repeater students was not incorporated in the analysis presented here nor in proceeding paper sections.

is significant enough to reject the null hypothesis at a p-value of 0.05, all of the observations for that entire grade in that year are removed from the dataset.

For each possible observed characteristic – previous year's test score, previous year's behavior grade, free lunch status, English language learner, special education, or gender – the one-way ANOVA implemented here tests whether or not the classrooms in a particular grade differ on the distribution of each of the six characteristics. This process is repeated on the full dataset separately for each of six traits, and entire grades for a school-year are removed only after the analysis of all six characteristics is complete. Table 6 shows the percentage of grades removed from the dataset by characteristic. The percentages of what this study assumes to be "tracked" grades per year generally align with Ammermueller and Pischke (2006), who find tracking in European elementary schools to be almost negligible. The list of the grade-school-years removed from the sample in this study is listed in Appendix B.

There are two noteworthy observations from Table 6. First, the percentages of unevenly distributed (or "tracked") grades in a given year remain consistent as the cohort moves through time. For instance, in the chart of English Language Learners tracked grades, the percentage of ELL tracked classrooms in grade 3 of year 1997 is generally in-line with percentage tracked in grade 4 in 1998. As such, the grades that are tracked over time remain consistent. That being said, although the percentage of tracked grades is consistent over time, there are more instances of tracking in 4th grade when compared for

2^{nd} grade. That is, deliberate tracking is likely to be more relevant as students age: this is logical as a school administrator will have much more historical information about a 4^{th} grader compared to a 2^{nd} grader, and as such, it is much more possible to separate out students as they proceed through school.

A second step in this identification method is to limit the trimmed sample strictly to those without special needs (Hanushek et al., 2003). A special needs student is defined as being a behavior problem or having ELL or special education status. From this trimmed sample, two sub-groups based on socioeconomic status are specified in order to identify differential peer effects between these two groups. The first group is composed strictly of all students receiving free lunch. The second sample includes all other students in the trimmed sample. Even though these two types of students are in the same classroom in Philadelphia, distinguishing between them allows for conclusions to be drawn regarding differential peer effects for free lunch and non free lunch students.

Table 7 describes the overall trimmed sample and free lunch and non free lunch sub-samples to be employed in this paper. By removing the special-needs students and classrooms of unequal distributions of student traits, this study has discarded the tails of the distribution of classrooms, and what remains are various samples of students in evenly assigned rooms. From these samples, the identification of peer effects can be obtained. Discarding the high variation created by principals in non-random assignment has generated a more challenging statistical task because the variation is diminished between grades. And yet, this is the correct variation of classrooms in which to evaluate peer effects, as self-selected biases are removed.

Strategy 2: Late Arrival

The second method of identification assesses the peer effects of a late arriving student into a classroom and relies upon within-classroom variation in peer group characteristics. As a result, this strategy can explore exogenous variation, or small unexpected perturbations, in classroom composition that is beyond the control of school management. This exogenous variation is sufficient for a quasi-experimental method.

Most generally, a late student, who arrives in school k in academic year t, will spend a minimum of 20 school days (i.e., one month) and a maximum of 160 school days (i.e., 8 months) in a given school. This student was not in school k in year $t-1$, thereby removing the possibility that school management has in-house records on this student at school k. Table 8 provides descriptive statistics on late arrivals and their non-late classmates who received a late arrival in their rooms. These statistics are based on the relevant population from which the analytical sample is derived. This population includes only those students in grades 3 and 4 and for whom the data contains a SAT9 reading or math test score.

To test for the random assignment of late arrivals, the analysis follows Sacerdote (2000). Employing a logistic regression model in which a binary variable is assigned a value of 1 if a late arrival possesses a particular observed characteristic, this test is designed to determine if a late arrival is matched according to the observable characteristics of the non-late students in his or her classroom assignment. If late arrivals are randomly assigned, this would yield evidence that all observable and unobservable similarities between late and non-late students would be no larger than what would be expected by chance.

The model implemented here is more complex than Sacerdote's (2000), however, because students can be assigned based on an array of observable characteristics; Sacerdote evaluates freshman dorm assignments as based only on previous academic scores. For elementary students, the process of matching late arrivals to their teachers and classrooms can be made on a multitude of student, neighborhood, teacher, or peer group characteristics. This study will assume that late arriving students will be matched on the basis of the same characteristics that were used to create the trimmed sample in the first empirical strategy: previous test performance, special education, ELL, behavior problem, free lunch, and female.

To conceptualize student matching, this paper presents what will be termed a 'resource-based' match. In this hypothesized assignment algorithm, it is assumed that special education students require more school resources than an ELL student who will require more school resources than a behavior problem, and so forth. In this framework, a principal assigns students based on the amount of school resources they need, from greatest to least. If a principal assigns a late arrival non-randomly, then, the principal would be hypothetically using this process described, in Illustration 1.

Implementing logistic regression, the non-random assignment process is evaluated for late arriving students. If the matching process is indeed utilized, then coefficients on the classroom characteristics will be significant. However, Table 9a shows that conditional on the non-late student characteristics of classroom *j* in school *k*, there is no relationship between late arrival *i*'s observable characteristics and the observable characteristics of the classroom to which he or she is assigned.

In more detail, each column of Table 9a is a node on the assignment tree from Illustration 1. For instance, in column 1, the dependent variable asks if the late student is a special education student, with a 1 as yes and 0 as no. It is also possible that the late arrival may possess other observable characteristics, though regardless he or she will certainly have special education status in this first column. As such, the independent variables in column 1's regression control for all other possible observable "down-branch" classroom characteristics (i.e., not simply the classroom percentage of special education students) that the student may embody. The models also control for teacher characteristics and school, year, and school-year fixed effects.

The results from column 1 demonstrate that there is no significant relationship between any observable characteristic of a special education late arrival and the observable characteristics of the classroom. That is, there is no association between the percentage of special education students in a classroom and the assignment of a late special education student to that same room. Nor is there any other relationship between other observable characteristics of the classroom and the observable characteristics of a special education late arrival. The only significant factor is that the late student is assigned to the classroom in that grade that has the minimum number of students.[8]

In column 2, the principal now assigns those late arrivals at the second node of the resource-based tree in Illustration 1. At this juncture, all special education late arrivals have been assigned to their classrooms, and the next type of late arrival requiring the second highest level of school resources is an ELL student, who again may or may not also possess other "down-branch" observable characteristics. However, as column 2 shows – and in fact all subsequent columns in this table show – there is no significant relationship between any classroom characteristic and any observable late arrival characteristic. In fact, the only driving factor is that the student is placed in the classroom with the smallest head-count for that particular grade.

Table 9b provides a slight iteration to the matching process, although the results are consistent to those of Table 9a. Because some late arrivals in school *k* in year *t* were in the Philadelphia School District in year *t-1* in a school other than *k*, it is possible to assess whether or not late arrivals can be assigned based on previous test scores that would have

[8] The variance in the head count in classrooms for a particular grade in a particular school is extremely small in this sample. In other words, out of three rooms in a particular grade, two may have 26 students and the third will have 25 students.

been recorded in the district's system. To do so, the analysis is analogous to Illustration 1. In this new iteration of the resource-based assignment process, it is assumed that special education students require the most school resources of all potential late arrivals. However, at the second node of the resource-based assignment process is the introduction of low and high scoring late students, as deemed by previous year's SAT9 reading score. Following the assignment of lower and high scorers are ELL students, behavior problems, and females.

With this modified matching algorithm, a similar analytical process is employed as before. The independent variable is binary, indicating whether or not a late arrival embodies a specific, observable characteristic at any particular point in the hypothetical classroom matching process. The independent variables are observable classroom characteristics, all in percentage form, of the classroom in which a late student was placed upon entry into school *k*. The results in Table 9b provide consistent results to those in Table 9a. The addition of low and high scorers in columns 2 and 3 of this table do not alter the conclusions: the observable characteristics of the classroom are not significantly significant in predicting late student placement. Instead, as with the results in Table 9a, the coefficients on being placed in the minimum class for the student's grade are the only significant determinants of classroom placement.

The conclusions from this analysis point to the fact that late arrivals are being placed according to minimum class size instead of by trait, hence yielding evidence of random assignment on observable characteristics.[9] That is, neither a late student's own characteristic nor the percent of the class that carries those traits imply anything about the matching process. This is made evident by the lack of significance in any of the matching coefficients, running down the diagonal of the chart.

Results: The Effects of Peers

A Baseline Model

Before integrating the identification strategies outlined in the previous section, this study first specifies the components of a benchmark education production function. Recall the value added production function developed in a previous section. Equation (6) had put-forth the following:

$$a_{ijkt} = \beta_0 + \beta_1 F_{it} + \beta_2 N_{it} + \beta_3 T_{jtk} + \beta_4 C_{jt} + \beta_5 P_{-ijkt} + \beta_6 a_{ijk(t-1)} + \gamma_{ijtk}$$

where academic year's achievement is a function of contemporaneous family, neighborhood,[10] teacher, classroom, and peer inputs as well as the previous year's achievement and an error term of school, year, and school-by-year fixed effects, individual shocks that vary over time, and a class-specific random component that is common to all members of the same class. This latter term is specified empirically as

[9] Note that class size is not highly correlated with observable classroom characteristics. The correlation coefficients between class size and the five observable characteristics from Table 9a range from 0.008 to 0.09. Furthermore, non-late' peer classroom characteristics were regressed on teacher characteristics and class size. The coefficient on class size is not significant. Results are available upon request.

[10] In the absence of family information, the vector of neighborhood variables will also serve as proxies for family data for each student.

robust standard errors adjusted for clustering within classrooms. Note that empirically, this model also controls for student demographic characteristics.

This value added model enables the estimation of specific components of educational production.[11] Using the analytical sample of all students, Table 10 presents two sets of least squares regressions using the SAT9 reading and mathematics scores as dependent variables. Both regressions contain robust standard errors clustered at the classroom level, and contain school, year, and school-by-year fixed effects. The mean of the reading dependent variable is 42.12 (*SD* = 14.65) and for math is 56.26 (*SD* = 19.01).

From these baseline results, there are several findings of interest among the entire span of covariates. First, the coefficients on student's own gender, race, English language learner, and special education are statistically significant and in the hypothesized direction (Argys, Rees, & Brewer, 1996; Caldas & Bankston III, 1997; Coates, 2003; Ogbu, 1989; Summers & Wolfe, 1977). In addition, students who have repeated the current grade have a higher reading score in this academic year whereas being young for the grade has no significant relationship with achievement. Lagged behavior grades are significant and positive – the higher last year's behavior grade, the higher this year's achievement score. Finally, having gone to kindergarten in the Philadelphia system is negative, although insignificant. This result may be explained by DeCicca (2007), who

[11] As a first specification check of the empirical value added model, current outcomes were regressed on future inputs (future inputs should not be correlated with current test scores). Notably, the coefficients on future peer inputs are not statistically significant, hence yielding evidence that the model is specified correctly.

suggested that the short-run positive impact of kindergarten depreciates considerably even by the end of first grade.

Neighborhood and teacher coefficients generally do not provide significant effects on achievement. Once being a free lunch student is accounted-for, for which the coefficient is significant and negative, neighborhood characteristics are not statistically significant in determining student achievement. Rather, it is the student's individual characteristics and school environment that affects achievement outcomes, not neighborhoods.

As for teacher characteristics, the lack of consistent significance aligns with many education production studies, including Hanushek (1989), Argys et al. (1996), and Nye, Konstantopoulos, and Hedges (2004). Cumulatively, these studies demonstrate that teacher gender, race, and experience do not significantly or meaningfully relate to student performance. The only exception is teacher race as other, which includes demographic populations such as Native Americans. The coefficient has a negative statistically significant influence on learning. However, the sample has a limited number of teachers with race as other, and hence this small sample can be driving the results. In addition, a teacher's education background does not make a difference in student achievement (Clotfelter, Ladd, & Vigdor, 2006, 2007; Goldhaber & Berwer, 1997; Hanushek 2006). Consistent with Hanushek (1992), class size reported in the table is positive, yet insignificant. The positive coefficient may imply that, for unobservable reasons, better teachers are assigned to bigger classrooms. In essence, class size may be picking up an omitted teacher effect. Consistent with Henderson et al. (1978) class size-squared is not significant.

Important in the baseline model are the coefficients on peer effects, beginning with peer academic ability as measured by: mean class ability, mean class ability squared, and the interaction between mean class ability and a student's own individual lagged test score.[12] Mean class ability for classroom *j* is defined as the average of the lagged test scores for all students in the classroom. Lagged test scores rather than current scores are utilized to calculate the classroom mean to avoid issues of simultaneity between individual and peer effects (Hanushek et al., 2003). The average achievement value for student *i* in classroom *j* always excludes a student's *i*'s own lagged achievement for each student. As such, there is a slightly different mean value for every student with a given classroom.[13] Both mean class ability and mean class ability squared are included in each regression.

Mean class ability is negative and statistically significant for both reading and math. Mean class ability squared is positive and significant for reading and math. On its own, it would at first glance seem that mean classroom ability is associated with lower student achievement. However, it is necessary to take both mean ability and mean ability

[12] The variability in classroom test performance was tested in this model and subsequent models in this paper. However, here and in the following models (based on differing samples), the coefficient on classroom test performance variance was consistently insignificant. Hence, it was not included in the regressions presented in this paper.
[13] As a test of robustness, the test scores of only those students who were not in the classroom the year before were used to comprise the mean classroom scores in reading and math. The results, though slightly larger, are consistent in sign and magnitude to those coefficients pertaining to classroom average ability, its squared term, and its interaction to individual test scores that are presented in this paper.

squared into account. Interpreted simultaneously and holding all else constant, these two results imply that, for reading and math, increasing average classroom ability beyond a specific threshold value has positive effects on individual-level student achievement. On the other hand, before average classroom ability reaches this threshold value, it is possible that lowering the mean classroom ability may also increase individual-level student achievement. It may be hypothesized that lower ability students may feel less intimidated around other lower ability students or that they may fall behind in classrooms in which the average pace of learning is higher than what they can handle.

The variable mean class ability interacted with lagged individual test scores is negative and significant for reading and positive and significant for math. For reading, holding all else constant, the negative interaction term indicates that raising the average classroom ability has its greatest benefits for students at lower ability levels. For math, this interaction term implies that raising the average ability of one's classmates has larger payoffs to students whose individual abilities increase. Thus, from the initial analyses of these baseline models, there is evidence student testing performance plays out differently in the classroom depending on the academic subject.

Indicators of other observable peer effects are in terms of classroom head counts. The peer variables are constructed in levels rather than as percents because of the small variance in class size for a given grade. Thus, having levels provides for a more precise estimation of the peer effects. The count of students receiving free lunch, which serves as an indicator of the number of poverty students in a room, is negative and significant and consistent with previous literature on peer effects (Hanushek et al., 2003). The coefficient on the count of behavior problems is also negative and significant, indicating that misbehaved students detract from classroom learning (Figlio, 2005; Lazear, 2001). The coefficients on the class counts of special education students and English language learners are negative though insignificant. Finally, an increase in the number of females in a classroom raises individual student achievement, as the coefficient is positive and statistically significant.

Implementing Strategy 1: Peer Effects in the Trimmed Sample

The results from the baseline model indicate the existence of a diversity of statistically significant peer effects, both in terms of classroom academic ability as well as in the count of students with specific observable characteristics. This section implements the first identification strategy as described previously. Doing so allows for the first specification of unbiased coefficients on peer effects, and subsequent evaluation of two student populations within the Philadelphia School District: free lunch and non free lunch students.

Table 11 provides selected results. For the sake of clarity, however, only peer results are presented. Nonetheless, each model contains all variables from Table 10 as well as school, year, and school-by-year fixed effects and robust standard errors clustered at the classroom level.

The peer effects results are generally similar to those in the baseline model. This indicates that either tracking is a relatively rare policy in elementary grades or that those grades whose classrooms that do have unequal distributions of student traits do not significantly impact the measurement of peer effects. As such, there is not a large amount of bias in the peer estimates in Table 10 once examining Table 11. In terms of classroom ability, the results remain as statistically significant in both reading and math models. The magnitudes of the coefficients here are slightly larger, indicating a stronger peer effect of classroom ability than the baseline model would have suggested. Mean classroom ability squared and the interaction between mean ability and lagged student ability are almost identical to the previous set of analyses. The difference, then, between this analysis and the previous lies in the strengthening of the coefficients' sizes on mean ability for both test subjects. Nonetheless, the interpretation remains the same as previously determined: there is statistical evidence that classroom ability has a meaningful and differential effect on achievement depending on reading or math. The effect depends on the test subject's classroom mean ability and its interaction with student lagged test achievement level.

As the observations in this table are narrowed based on the equal distribution of student characteristics by grades, the coefficients are slightly more negative in size and may reflect more unbiased measures of peer effects than those provided by the baseline model. Larger negative coefficients on the classroom average ability for both reading and math (with a similar squared term to what was seen previously) indicate a larger threshold value than before at which mean classroom ability would positively affect individual student test scores. That is, it takes a higher mean classroom ability level to positively influence individual student achievement than then non-random sample would have indicated.

Observable peer traits are in-line as well with previous results. The classroom counts of free lunch students and behavior problems negatively and significantly affect student achievement in both reading and math, whereas the classroom count of females positively affects achievement in both test subjects. A similar positive and negative spillover interpretation can be told here as before. Being a free lunch recipient or a behavior problem leads to lower academic achievement, and this is exacerbated when students with similar characteristics. On the other hand, females provide positive spillover effects in the classroom, and can potentially offset the negative achievement effects of their peers.

In the economics of education literature, the effect size is most commonly defined as the standardized regression coefficient (Ammermueller & Pischke, 2006; Hoxby, 2000; McEwan, 2003). For the sake of comparability to other literature on peer effects, Appendix C presents the effect sizes of the observable student characteristics for all models and samples going forward. The results are generally consistent with the effect sizes of other studies in peer effects (Ammermueller & Pischke, 2006).

Implementing Strategy 2: Late Arrivals

Using the identification strategy of assessing the peer effects of late arriving students in a classroom allows for a second method of obtaining unbiased peer effects estimates while simultaneously confirming the results from the above analyses. Importantly, this strategy provides a sense of the marginal classroom peer effect, as a late arriving student is the final student to be placed in a classroom in a given year. As such, the late arrivals exogenously (and as proven in a previous section on identification, randomly so) disrupt the classroom composition, thereby providing a quasi-experimental method for assessing the effects of peers. As a result, the estimates here can contribute to the overall goodness-of-fit of other models used in the study.

In this strategy, the evaluation relies on the trimmed sample, with only a slight difference made here. The distinction between the trimmed sample in strategy 1 and the sample of students evaluated in this section is that the observations of the late arrivals have been removed from the trimmed sample. This was done specifically to examine the effects of late arrivals on their non-late classmates in classroom *j*. Tables 12a and 12b provide estimates of three different strategies that evaluate the effect of late arrivals on the reading and math performance of other, non-late students in the room. In these tables, only the results of observable characteristics of the late arriving students are provided, for the sake of clarity. Except for those variables pertaining to the lagged test score of the late arrival, the characteristics of the late students are binary variables, which equal to 1 if classroom *j* received a late student embodying a particular characteristic displayed in the table.

In strategy A, only those classrooms who received late arrivals without any missing data on the six observable characteristics were evaluated (i.e., this explains the much smaller sample size in the first columns of Tables 12a and 12b).[14,15] Note that the column of results for strategy A presents four different regression models simultaneously. In each model, non-late student-level reading (Table 12a) or math (Table 12b) scores were regressed on a late arrival's lagged reading score (Table 12a) or math score (Table 12b), the late arrival's score interacted with the non-late student's own lagged test score for the given subject, an indicator signaling if the student was a behavior problem,[16] as well as the inclusion of one other indictor of the late arrival – either free lunch, ELL, special education, or gender. Each of four indicators was included in a separate model. Aside from school fixed effects and classroom clustering, no other covariates were included in these models.

Thus, for the column of strategy A results, the coefficients for the late student being free lunch, ELL, special education, and female are all from four separate regression models. The three additional variables presented in both tables – the lagged test score of the late

[14] Approximately 90-percent of late arrivals do not have lagged information as they came from outside of the District. However, some students did arrive from other schools in the District, and thus there is lagged information available to create a full vector of non-missing data for these students.

[15] Note that the observable descriptive characteristics for late arrivals with and without missing information are similar. This table is available upon request.

[16] Being a behavior problem was determined analogously to previous baseline and trimmed sample analyses.

student, the interaction with the non-late student's test score, and late being a behavior problem – are from the model which included an indicator for the late student holding free lunch status. However, there is a consistency in the estimates of these three variables across all models.

The peer effects of free lunch, ELL, special education, and female are fairly consistent with previous baseline and trimmed sample analyses. There is a negative effect of late arrivals with free lunch status that is similar in magnitude and statistical significance to the estimates in Table 11. Also similar are the effects of females. Having a female late student in the classroom impacts reading and math performance consistently across subjects in Tables 12a and 12b and across analyses presented in Table 11. Although special education and ELL status are statistically significant in these analyses, they lose their significance in strategies B and C to follow. The sudden significance of these two characteristics in strategy A are not worrisome, however, because the model selected in this evaluation is slightly different, and the sample is much smaller. Nonetheless, strategy A provides a first indication in the robustness of the peer effects evaluated in both baseline and trimmed samples.

The econometric specifications in strategies B and C are similar in construction to previous models in Table 11. Including the late variables presented here, all other covariates from the original regression equations have been included in the model. Note that student, teacher, and neighborhood characteristics are exactly the same as those used in all regressions for each student *i* in classroom *j*. However, to avoid multicollinearity issues when adding the late arrival into the regression, class size and peer classroom count variables in classroom *j* have been altered from previous regressions. Each variable has been split into a late arrival identifier and non-late student count for each classroom *j*, though each corresponding pair sums to the original values. For example, for the classroom count of free lunch students, these regressions have a variable for the non-late number of free lunch students in the classroom and a variable (binary) for receiving a free lunch student. Together, "non-late free lunch class head count" and "late free lunch class head count" will equal to the classroom value of the total number of free lunch students in classroom *j*. It is possible that a classroom does not receive a late student assignment in the school year. As such, non-late class head count will identical to the original class head count for a given variable, and late arrival indicator will be 0.

Strategy B and C can be viewed as complements for testing the robustness of the peer effects in Table 11. Strategy B first predicts lagged reading (Table 12a) and math (Table 12b) achievement and behavior information for all late arrivals based on the sub-sample of late students. From this sub-sample, the significant predictors of lagged test scores were used to predict the test scores for those late arrivals who did not have lagged test score information. A similar process was conducted to predict lagged behavior scores. From this, the model in column B was run for the non-late students in the trimmed sample in which all late arrivals had full information. On the other hand, strategy C implemented a model in which lagged information was not incorporated into the regression, neither for late arrivals nor for non-late students in the sample.

The peer effects for strategies B and C are consistent with the results from strategy A and the previous set of analyses in Table 11. This consistency is apparent across both testing subjects. First, the late arrival's lagged testing information is not a significant predictor. This is logical, however, as the late student is only contributing approximately $1/26^{th}$ to the mean average ability for the entire classroom. In essence, the average ability of one student's entrance into the room does not alter the mean in any meaningful capacity. Hence, the lack of significance on lagged testing measure for late students is unsurprising.

The robustness of the late arrival strategies B and C is apparent, however, in the evaluation of observable characteristics. Because the peer effect variables are binary for the observable characteristics, the effect sizes are fairly straightforward from the coefficients presented in Tables 12a and 12b. The coefficients on free lunch status are negative and significant and fall in-line with the baseline and trimmed estimates. In this case, the effect of a student on free lunch status is associated with a 0.11 to 0.14 decrease in reading achievement and a drop of 0.10 to 0.19 in math for other students in the class. Behavior problems continue to be significant and negative. Adding a randomly assigned late behavior problem to a classroom will decrease average test performance by 0.23 points for reading and 0.22 (strategy B) in math for other students in the class. Finally, females boost academic achievement. As before, the peer effect of a female can almost entirely offset the negative peer effects of a behavior problem, holding all else constant.

The results from these three strategies demonstrate the robustness of the coefficients from the baseline and trimmed sample analyses.[17] In the trimmed strategy, the case was made that the sample had been de-tracked, retaining only those classrooms-per-grade per school-year that did not have statistically significant differences in observable characteristics. Here, the case was made that late students are randomly assigned to classrooms: observable characteristics cannot foretell the placement of a late student and only class size matters. Both strategies intended to implement quasi-experimental methods, and both present similar coefficients of peer effects. It suggests non-linearities in changing classroom composition.

In addition to confirming the results of the trimmed sample analyses, the results of this second strategy also contribute an additional concept used in managing the effect of peers: the marginal peer effect. Since the last student to enter the classroom is this randomly-assigned late student, the analysis in this section provides insight into the effects of altering classroom composition. This is where this evaluation turns next.

---

[17] As a further test of robustness, regressions were run to determine if classroom attributes in turn affect the test scores of late students. The regressions are similar in form to those of strategy 1. The results show that classroom peer effects (in terms of average ability and counts of observable classroom characteristics) are significant. Thus, peer effects may be influential on this late student, just as he or she in turn may affect the room.

Policy Analysis

Given the consistency across all analyses thus far and in conjunction with the linear relationship of adding classroom peers, this section examines how changing peer groups can alter educational outcomes. A more detailed examination of how the effects of peers can impact testing performance involves partitioning the trimmed sample from strategy 1 into two subgroups: free lunch recipients and non-recipients. Assigning all students in the trimmed sample to either group allows for the evaluation of how peer groups play out differently in an urban school district depending on socioeconomic status, controlling for other student, classroom, teacher, and neighborhood characteristics.

Status Quo Peer Composition

To begin, Table 13 provides the peer effects results for the trimmed sample broken out by free lunch and non free lunch students, both of whom are non-special needs students. The coefficients on the three measures of classroom ability are consistent with previous analyses: negative, highly significant coefficients on mean class test performance, positive (generally significant) coefficients on mean test performance squared, and statistically significant negative coefficients on the interaction for reading and positive, insignificant coefficients for math.

Focusing on other observable characteristics, peer effects play out differently for free lunch students than for non free lunch students. The results indicate that free lunch peers negatively and significantly affect other free lunch students' reading achievement by 0.15

points per count of free lunch students in the room. Math achievement for free lunch recipients, however, is not significantly impacted by the count of free lunch students in the classroom. On the other hand, free lunch peers do not statistically significantly affect reading achievement of non free lunch students, though mathematics achievement for non free lunch students is statistically significantly decreased by 0.22 points per free lunch student in the room in math. On average, a non free lunch student is in a classroom of approximately 8 free lunch peers in the district.

The classroom count of behavior problems consistently, statistically, and negatively affects only free lunch students by 0.25 points per student count in reading and 0.50 points per student count in math. On the other hand, behavior problems do not significantly impact the achievement of non free lunch students. This might imply that free lunch students are more at risk when in classrooms with behavior problems, whereas non free lunch students are more resilient to the behavior problem composition of their classrooms. Finally, the count of females positively affects classroom learning in three out of four regression models, indicating that as consistent with the previous sets of analyses, the count of females may offset the negative spillovers of behavior problems or free lunch students, depending on the sample and academic test subject.

Altering Classroom Composition

This study next evaluated the effect of increasing (or decreasing) classroom counts of particular groups of students. Because the coefficients from the results section were similar for the average peer effect (i.e., trimmed sample strategy) and the marginal peer effect (i.e., late arrival strategy), the analysis allows for the manipulation of peer groups in the classroom without having to address non-linear relationships. This study will manipulate classroom composition for two separate analyses: within-district and within-school.

Within-District Classroom Changes. To begin first with an evaluation of within-school classroom variation, this study examines the effects of changing peer groups across the entire district of Philadelphia. That is, this section asks if it is feasible to rely on the district as a whole to impact learning. If possible, then doing so provides insight on how district administrators could utilize the variation across their entire domain of schools in making decisions regarding improving achievement through different peer groupings. In other words, this section asks if principals can implement changes to student learning by relying on district resources (i.e., not simply single institutional resources).

To begin, Table 14 provides the impact of altering peer groups on additional months of learning in a single school year, based on the regression coefficients from Tables 11 and 13.[18] This is accomplished by increasing the district average classroom head-count by one standard deviation for a given characteristic.[19] This is done for each observable student trait and conducted separately for each. Note that the standard deviation in each

[18] The conversions from coefficients into months of learning are based on May and Supovitz (2006).
[19] This experiment, of course, is hypothetical. It is not possible to increase each head count by one standard deviation: a classroom who receives 1 standard deviation more girls will have to place the boys in a different classroom, thereby lowering the headcount of girls in that other room.

characteristic is determined across all classrooms within the district. In other words, this is the variation across the entire trimmed sample, and broken out by free lunch versus non-free lunch status.

Table 14 shows that it is possible to increase months of learning in a single school-year by changing classroom composition across the district. For example, it is possible to increase the number of free lunch students in a classroom by one standard deviation. Students across the entire Philadelphia School District are assigned to classrooms with average counts of approximately 10 free lunch students and a standard deviation of about 9 students. By changing the classroom count from 10 to 19 free lunch students, this process has in essence transforming a South Philadelphia classroom into a West Philadelphia classroom. Thus, the results here, based on the previous trimmed analysis regression coefficients, imply that that in a given school year, there would be a decrease (increase) of a half to full month's of learning in reading per standard deviation increase (decrease) of free lunch recipients in the classroom. The precise gain or loss depends on whether the sample in question valuates the full trimmed or that broken out by SES across the district.

Moreover, in some inner-city classrooms, if it were possible to move a significant portion of free lunch students to the districts of neighboring suburbs (e.g., busing out 3 standard deviations of free lunch students) then the remaining free lunch students in the district would gain 1.5 months of learning in reading in that year. For math, non free lunch students would gain almost three months of learning. This result for the peer effect of moving free lunch recipients in and out of classrooms, among other peer effects, suggests that peers do matter, not only in current classroom composition, but also for policy purposes in the way that a school district can change the learning environment.

Over many years of schooling, it is clear that this result can build up to sizeable learning gains. To elaborate this point, the analysis evaluates the process of dynamic learning (Wolpin & Todd, 2003), by which the educational experience in one year of schooling impacts future learning. To evaluate the dynamic learning outcomes via a change in peer groups, this study examines two hypothetical processes over a three year period: initial impact and continual impact. In the initial impact scenario, there is a one-time initial shock to the classroom environment: a peer group increases by one standard deviation from the district average in the first year – holding all else constant. However, in the following two years, the classroom composition returns to the district average. Nonetheless, the change in peer group from the first year impacts individual-test scores in that same year, thereby also changing the lagged test scores to be evaluated in the next year, and the next, and so forth. In the continual impact scenario, the peer group increases from the district average by one standard deviation in each year. This not only continually impacts the current test score in every year but also those lagged test scores to then be evaluated in a subsequent year. Again, all else except for the change in peer group is held constant.

Tables 15 and 16 present the results of initial and continual impacts, respectively, for those students in the full trimmed sample and broken out by SES sub-samples. The results are presented in terms of months of learning, much like Table 14. In these two tables, the results provide the cumulative impact on months of learning after three years in elementary school. For Table 15, this implies the effects on learning over three years of elementary schooling from changing the peer group composition strictly in the first year. Table 16 presents the cumulative effects on learning over three years, when a particular classroom composition has been manipulated in all three years.

Beginning with initial impact learning, the results from Table 15 indicate, for instance, that increasing the number of free lunch students in an average district classroom by the one standard deviation in year one will decrease learning in reading by approximately 1.5 months over the entire span of three years of learning for the free lunch students in the room. In other words, a single year's manipulation of the classroom count of at-risk peers has been detrimental on reading performance over time. Also in reading, free lunch students experience decreased learning from an increase in number of behavior problems and decrease from class count of females. These results, however, suggest that free lunch peers have the largest impact on learning for free lunch students, holding all else constant. It is clear that differential results permeate throughout the table. Therefore, changing the classroom composition in the initial year can affect learning for years afterwards and for different students based on SES.

The results from Table 16 demonstrate similar interpretations on learning effects, though ones that have been exacerbated by the fact that the peer group has been altered three years in a row. For example, increasing the class count of free lunch students by one standard deviation in each of three years causes a decrease of 3 months of learning over time for free lunch students in the trimmed sample. As mentioned, this results not only from the impact of the peer group on learning the current year, but also because lagged learning impacts future learning (i.e., the test score lag in the value added models from the regression analyses presented previously). Looking from another perspective, decreasing the classroom count of behavior problems by one standard deviation each year actually increases learning by a little more than 2.5 months for free lunch students. Together this may indicate that if a classroom experiences an increase in free lunch students, it may be possible to offset the negative effects on free lunch students by decreasing the classroom count of behavior problems.

Within-School Classroom Changes. Rather than examining changes to peer groupings for the district as a whole, it is possible to look at altering classroom composition based on within-school variation. Doing so provides insight as to on how principals could influence classroom learning solely by relying on their institutional constraints. In other words, this section evaluates if it is possible to impact student learning simply from moving students around from within the same building.

As an example, this section examines three schools, each based on the percentage of free lunch students in their respective student bodies. They include: a school at the 25th percentile of the distribution of the percentage of free lunch students in the district, one at the 50th percentile (which is in essence the district average), and one at the 75th percentile. Each of these individual schools can essentially represent a different category of school and educational environments within the district.

Table 17 provides an example of the impact of altering one type of peer grouping within these particular schools: an increase in the number of free lunch students per classroom.[20] The results are in terms of month of learning in a single year on the same-year's months of learning, based on the regression coefficients from Tables 11 and 13. First, note that there is less variation in schools at the 25th and 75th percentiles of free lunch students, which have standard deviations of 4 free lunch students and 7 free lunch students, respectively. This indicates that as schools become more homogeneous in either direction (i.e., less or more free lunch students), there is less variation in classroom composition. This explains the higher standard deviation for a school at the 50th percentile, which in essence represents the district average.

Table 17 demonstrates similar results those in Table 14. There are negative, statistically significant effects of increasing the classroom count of free lunch students across all three

[20] As in the district-wide analysis, this scenario is purely hypothetical. It would not be possible to increase all classrooms in a grade by a single characteristic in a given school. However, this exercise provides an example of increasing a specific characteristic for any given classroom, without regards to the ripple effects on other rooms.

schools here, though the sizes of the learning effects differ by student free lunch status, test subject, and school poverty composition. For initial and continual impact analyses, Table 18 also presents similar results to Tables 15 and 16 with highest effects at the average school. Thus, Table 18 provides insight on the capacity of principals to increase (or decrease) learning within their respective schools over a three year period. Although there is a larger ability to institute changes to learning in schools in more heterogeneous schools, this particular analysis has shown that is it possible to impact learning in any school over multiple years of learning and for students of varying SES.

Overall, the results from both within-district and within-school analyses indicate that different policy objectives would need to address each peer effect differentially for many peer channels and for socioeconomic status. For example, there are mixed effects of the impact of the class count of free lunch students on other free lunch students, depending on the test subject. Thus, the way in which peers are reorganized has differential policy implications depending on if the district or any particular school desires to impact reading or math scores, or perhaps both simultaneously. On the other hand, it is entirely possible that moving behavior problems away from free lunch students diminishes the risk that they face academically in both reading and math. For non free lunch students, however, there is evidence of a lack of peer effect of behavior problems. Finally, a higher count of females in the classroom is beneficial to individual test performance, regardless of test subject or socioeconomic status. In essence, the peer effects of females may equalize other negative peer effects across the board.

Of course, there are a variety of objectives and even more constraints. It may appear that mixing students and exacerbating or mitigating a variety of peer effects can have policy implications for both free lunch and non free lunch students in the same urban school district. However, the outcome is dependent on what objective is chosen and what constraints are placed on moving students within and across schools.

The Effects of Peers on Non-Academic Outcomes

The analysis in this study has thus far focused on the effects of peers on testing performance as an educational outcome. However, before concluding, one final set of outcomes is briefly assessed. Specifically, this section examines the effects of classroom composition on non-academic outcomes – behavior grades and truancy rates.

Table 19 provides the regression results of these two outcomes on all three strategies employed in this paper: baseline, trimmed sample, and late arrivals. The first set of results is based on a logistic regression model. Here, the dependent variable is a binary indicator, determining if a student received an "A" or "B" in behavior in the current year *t*, controlling for all else including a lagged behavior score. The second set of results is based on ordinary least squares, in which the dependent variable is the rate of unexcused absences on a student's record in the current school year. This model incorporates a lagged measure of this truancy rate, as well as all previously employed covariates.

The results for behavioral outcomes demonstrate that the observable classroom environment has the capability to predict relatively "good" student behavior in the current year. Consistently across all three regressions, the classroom count of behavior problems and ELL students predict a decreased probability of receiving an "A" or "B" in behavior for year *t*. Having a higher classroom count of females increases the probability of receiving an "A" or "B" in the current school year, holding all else constant. Put another way, having a higher classroom count of females decreases behavioral risk, as would having lower counts of behavior problems and ELL students in this analysis. These results indicate that the classroom learning environment impacts both testing performance and behavioral outcomes.

The results differ for truancy as an outcome. Here, only an increase in free lunch or ELL students is indicative of higher truancy rates. The lack of significance on classroom attributes, however, puts forth a potentially accurate depiction of the causes of truancy. Rather than a function of the classroom environment, higher rates of truancy may arise from family environments (Kearney & Silverman, 1995; Sheldon, 2007). Furthermore, truancy increases as family and school SES decreases (Orfield, Losen, Wald, & Swanson, 2004; Swanson, 2004). Hence, the overall lack of significance of classroom factors, except for the measure of free lunch status and English language learners, perhaps sheds insight into those student-level factors that represent family environments and their effects on absences.

CONCLUSION

By developing a theoretical base of student achievement from the education production function and subsequently analyzing large-scale, longitudinal data of individual- and multi-level observations, this investigation has provided insight into the causal effect of peers and has thus contributed to the literature surrounding these issues. Having a unique and comprehensive dataset has facilitated two novel identification strategies, both of which have allowed for quasi-experimental methods to be employed. In emulating random assignment, these strategies together have surpassed previous endogeneity issues of self-selection and non-random classroom placement. The first strategy, implemented on a sample of equally distributed, non-special needs students, has provided estimates for peer effects of free lunch students and non free lunch students. The second strategy, which has identified late arriving students, confirms the results from the trimmed sample and provides insight into the marginal peer effect. This linear relationship of the effects of peers is what has allowed for a policy discussion on the academic consequences of moving students on gains and losses of months of learning.

Overall, this paper has presented evidence that there are significant peer effects in reading and math standardized test achievement, even after holding constant student and neighborhood demographics, teacher characteristics, and classroom attributes. Furthermore, even when controlling for a variety of channels of peer effects, other peer effects continue to surface as significant predictors of test performance. In addition, the effects of peers remain significant even under the empirical specifications of the value added model, in which the lagged test performance is assumed to soak up all historical information about an individual student. That is, above and beyond all current and historical attributes in a student's schooling environment, the effect of peers still remains significant.

In this study, peer effects surfaced across multiple domains, including academic characteristics, SES, and gender. Evaluating separately different peer characteristics for free lunch and non free lunch students demonstrated that free lunch students were more at-risk for negative aspects of classroom composition than were their higher SES counterparts. This held true among many observable peer characteristics across both testing subjects. As an example, free lunch students were significantly affected by behavior problems in both reading and math, whereas non free lunch students are not. This result arose as negative effects on months of learning not only in the current year, but also for several years to follow. This was demonstrated in both initial and continual impact analyses.

However, the peer effects were not materialized in the same way for all groups of students. For example, non free lunch students experienced an increase in testing performance (and months of learning) by a decrease in free lunch classroom peers in math, whereas the same is not true for the free lunch students in the classroom. Similarly, an increase in the classroom count of females improved testing performance and learning for non free lunch students in both reading and math; only in reading did an effect exist for the free lunch student sub-sample. This demonstrates that a large number of significant combinations exist as to how to improve testing and learning in the classroom, depending on the test subject and the student. As such, policy implications from peer effects are prevalent and significant though not clear cut.

One thing for certain, however, is that it is possible to increase elementary school learning with the strategy of shifting peers. The within-district and within-school analyses demonstrated that changing the classroom composition of a particular student characteristic by one standard deviation can lead to significant monthly gains, holding all else constant. In one year, gains were up to one-month large, depending on test subject and sample. Over 3 years of shifting peers, however, gains were almost 5 months large in some instances. Thus, what this policy implies is that there are improvements in learning for relatively no money spent, other than the none-to-small costs of moving students across rooms (or potentially across schools).

Other policies have experienced similar or fewer gains in monthly learning, but at much higher prices. May and Supovitz (2006) found that for Black students in the Rochester School district, the America's Choice program added approximately one-half a month of learning per year in grades one through three in reading and math. However, the set-up cost is approximately $90,000 per 30 teachers and $2000-$4000 in instructional materials per classroom. Greene (1998) found that ELL students had gains of 3 months over 2 years of bilingual education programs in California. However, this translates to approximately $2,000 per student, because these programs require supplemental instruction (Chambers & Parrish, 1992). Finally, Borman and his authors (2005) evaluated the Success for All literacy program. A majority of the monthly gains were slightly larger than one month of learning, but the authors acknowledged the extremely high costs of the program of, at minimum, $135,000 per school over the first three years of implementation.

While these other programs do show promise in narrowing the achievement gap, they come at a fairly hefty price. Ironically, the districts which most desperately need these programs are those districts that cannot afford to purchase them. What this suggests, then, is that resource-constrained school districts must turn to alternative solutions in improving learning. This study has offered one such alternative. The focus on a single urban school district has enabled this study to document patterns of peer effects as students progress through early years of their schooling experiences. The analysis in the present study has demonstrated that not only do peers matter in a given school year, but that they matters across multiple measures of achievement, and that they matters persistently.

Although the implications of this paper support that altering classroom composition along many lines can improve student performance, further research may suggest how to do so. For example, given the results from this paper on the monthly gains over three years of elementary school learning, future research may construct an optimization strategy to maximize learning outcomes based on the allocation of students. This optimization process, as first implemented by Arnott and Rowse (1987) would utilize an objective function, subject to a variety of constraints, and the learning technology defined in this paper by the education production function in order to determine an optimal allocation of students whose peer effects would maximize the sum of scores. Asking "optimization for whom" is a logical response, to which an answer based upon the results from this paper can provide a substantive foundation.

Illustration 1:

Resource-Based Assignment Process of Late Student (if assigned non-randomly)

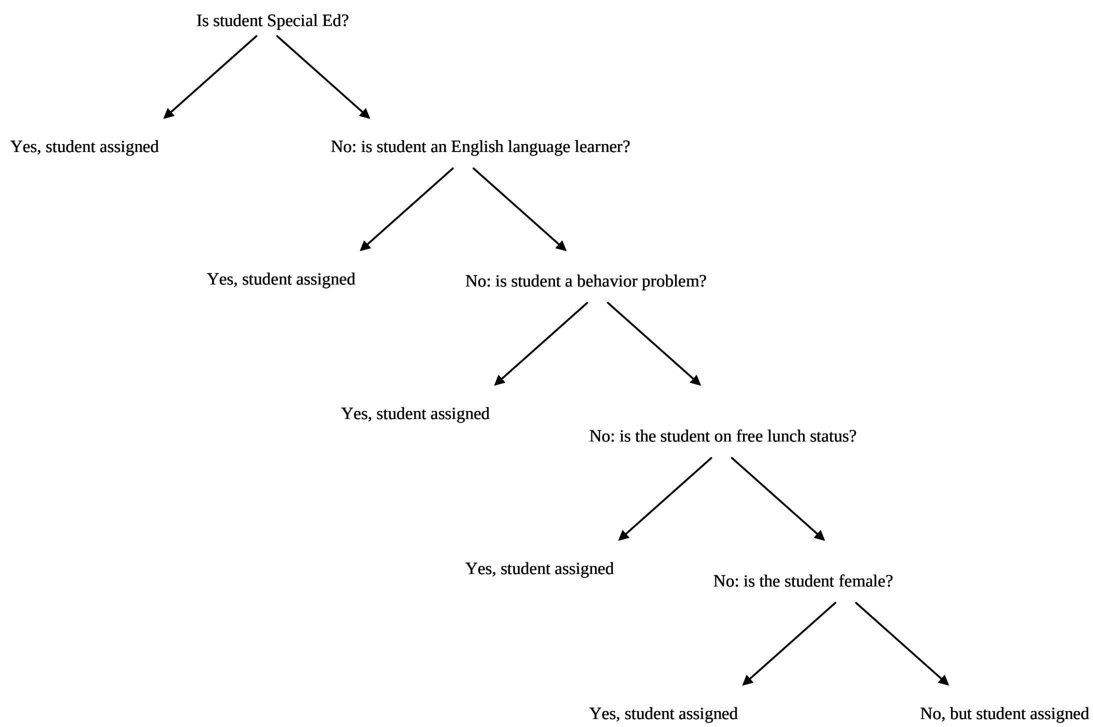

Table 1

Demographics of Urban Districts in the United States: 2008-2009 School Year

	% Minority	% White	% Elementary	% Free or Reduced Lunch	% English Language Learner	Graduation Rates
Baltimore	92.2	7.8	50.3	74.6	2.0	62.6
Boston	87.0	13.0	46.3	62.0	11.0	57.9
Chicago	91.1	8.9	52.4	84.3	13.3	54.3
Cleveland	84.6	15.4	52.6	60.0	9.2	60.1
Philadelphia	**86.7**	**13.3**	**47.4**	**84.6**	**5.7**	**62.0**
Pittsburgh	64.7	35.3	55.0	69.0	36.0	64.0
Washington, DC	93.0	7.0	43.8	70.0	9.3	68.0
Average	**85.6**	**14.4**	**49.7**	**72.1**	**12.4**	**60.7**

Table 2

Summary of Peer Effects Studies

Study	Peer Measure	Effect Size
Summers and Wolfe (1977)	Academic ability	Not available
Henderson, Miezowski, & Sauvageau (1978)	Academic ability	Not available
Evans, Oates, & Schwab (1992)	Behavior	0.00
Hoxby (2000)	Academic ability	0.40
Zimmer & Toma (2000)	Academic ability	0.04
Sacerdote (2001)	Academic ability	0.06
Cullen, Jacob, & Levitt (2003)	Academic ability	0.00
Hanushek et al. (2003)	Academic ability; SES	0.05
McEwan (2003)	Mother's education	0.27
Angrist & Lang (2004)	Academic ability	0.00
Ammermueller & Pischke (2006)	Number of books at home	0.11
Neidell & Waldfogel (2008)	Preschool attendance	0.01

Table 3

Student Panel Data Coverage by Cohort, Grade, and Year

	Academic Years							
Cohort	**1994/1995**	**1995/1996**	**1996/1997**	**1997/1998**	**1998/1999**	**1999/2000**	**2000/2001**	**2001/2002**
A	2	3	4	5	6	7	8	
B	1	2	3	4	5	6	7	8
C	K	1	2	3	4	5	6	7
D		K	1	2	3	4	5	6
E			K	1	2	3	4	5

Table 4

Descriptive Statistics of Full Dataset

	Population*		Analytical Sample	
	Mean	***SD***	**Mean**	***SD***
N	97,007		34,450	
SAT9 achievement outcomes				
Reading	39.11	*15.36*	42.12	*14.66*
Math	55.90	*19.00*	56.26	*19.01*
Reading, lagged	37.13	*15.57*	37.74	*15.49*
Math, lagged	56.92	*18.59*	56.99	*18.14*
Race, in percent				
White	16.82	*37.41*	17.67	*38.14*
Black	67.85	*46.70*	67.69	*46.77*
Hispanic	10.99	*31.28*	10.43	*30.57*
Asian	4.17	*19.98*	4.07	*19.76*
Other	0.16	*4.02*	0.14	*3.69*
Gender, in percent				
Male	48.90	*49.99*	48.86	*49.99*
Female	51.10	*49.99*	51.14	*49.99*
Academic variables, in percent				
Attended Phila kindergarten	85.14	*35.57*	84.70	*36.00*
Free lunch eligible	52.46	*49.94*	53.32	*49.89*
English language learner	3.65	*18.76*	3.18	*17.57*
Special education	4.04	*19.68*	3.34	*18.08*
Lagged behavior = D	10.54	*30.71*	10.92	*31.19*
Lagged behavior = C	22.83	*41.97*	23.49	*42.39*
Lagged behavior = B	35.29	*47.79*	35.05	*47.71*
Lagged behavior = A	31.34	*46.39*	30.54	*46.06*
Student's census block				
Block percentage: white	29.39	*32.50*	30.00	*32.87*
Block percentage: poverty	14.35	*8.67*	14.14	*8.67*
Block percentage: house vacancy	12.95	*9.38*	12.74	*9.31*
Log of income (in dollars)	10.15	*0.45*	10.16	*0.45*
Teacher race, in percent				
White	82.74	*37.79*	80.58	*39.56*
Black	16.26	*36.90*	18.41	*38.75*
Hispanic	0.68	*8.21*	0.69	*8.25*
Asian	0.28	*5.26*	0.27	*5.19*
Other	0.05	*2.20*	0.06	*2.47*
Teacher gender, in percent				
Male	7.88	*26.94*	9.41	*29.19*
Female	92.12	*26.94*	90.59	*29.19*
Teacher skills				
Teacher experience (in years)	3.81	*7.70*	3.89	*7.73*
Teacher state certified (percent)	24.23	*24.23*	94.70	*22.40*
Teacher has a masters degree (percent)	33.84	*33.84*	14.33	*35.04*
Class size (head count)	28.23	*3.80*	28.79	*3.51*
Academic classroom characteristics				
Mean SAT9 reading score	33.61	*11.35*	36.35	*10.90*
Mean SAT9 math score	53.69	*12.46*	53.99	*13.84*
Other classroom characteristics (head count)				
Free lunch	12.62	*7.54*	10.46	*8.47*
Behavior problems	1.62	*1.75*	1.80	*1.74*
English language learners	1.33	*3.16*	1.29	*3.06*
Special education	1.03	*1.54*	1.07	*1.43*
Female	10.37	*3.62*	10.65	*3.61*

*Note: Population is based on having observations with required test scores. Analytical sample is based on test scores and non-missing information for required independent variables.

Table 5

Qualitative Evidence from Former Principals, Principals, and Teachers

Individual	Source	Insights
Monk (1987)	Interview	“As the principal put it: ‘(It’s) just very random, no real look at any criteria for the simple reason that sometimes at the elementary levels that’s the best kind of grouping” (p. 170).
Former Principal	Interview (2010)	"There are an equal number of behavior problems per classroom." "…most schools [in Philadelphia] organize heterogeneously, keeping equity in mind."
Teacher	Interview (2010)	"Classes are generally diverse with a variety of reading levels."
Former District Leader	Interview (2010)	"Almost all of the elementary schools I know of, whether suburban or urban, the principal does the scheduling."
Teacher	Interview (2010)	"There is a maximum number of students per room: it's based upon district policy. Once a class goes over the limit, it will be split (pretty randomly) into two separate classes."
Principal	Interview (2010)	"You have to look at the numbers because its class size that you're trying to maintain." "I do not group children homogenously." "There is no formula…there is no pattern." "There is no tracking: I do not group kids. I do not try to get a specific 'type' in a room." "You just try to balance boys and girls." "Within a single classroom, you're going to have a wide range of students' abilities. This doesn't matter however, because you pull kids out for separate reading time with the teacher anyways."
Principal	Interview (2010)	"Homogeneous groups are considered not good. This will create inequities. Instead, there has to be a full curve range of ability." "There's not a lot of thought in the process. You really just want to have an equal balance of boys and girls in each [room]." "A typical neighborhood school in the district is probably just looking for a gender balance."
Principal	Interview (2010)	"The push of the district is heterogeneous assignment." "You want a fair amount of learning to take place, so you would never group one type of student together." "You balance out girls and boys." "Behavior problems are not too much of a factor in classroom assignment."
Principal	Interview (2010)	"At such a young age in elementary school, you really just look to balance out girls and boys. This might change in middle or high school, but my elementary school rooms look all the same for each grade." "In elemetnary school, so many of these kids 'look' the same on paper, that its really difficult to tell them apart. So really, the only way I can do it is by gender: half boys and half girls, or as close as I can get."

Table 6

Percentages of Grades per Year Demonstrating Evidence of Unequal Student Distributions

ELL

Year / Grade	2	3	4
1995	4.7%	0.0%	0.0%
1996	0.0%	6.9%	6.4%
1997	0.0%	6.9%	6.3%
1998	0.0%	2.9%	6.3%
1999	0.0%	6.9%	8.0%
2000	0.0%	0.0%	8.0%

Special Eduation

Year / Grade	2	3	4
1995	0.6%	0.0%	0.0%
1996	0.0%	2.3%	2.3%
1997	0.0%	4.6%	5.7%
1998	0.0%	4.6%	9.1%
1999	0.0%	4.6%	9.8%
2000	0.0%	4.0%	15.4%

Free Lunch

Year / Grade	2	3	4
1995	8.1%	0.0%	0.0%
1996	0.0%	9.2%	6.4%
1997	0.0%	8.6%	8.5%
1998	0.0%	0.0%	0.0%
1999	0.0%	10.3%	4.0%
2000	0.0%	0.0%	6.3%

Behavior Problems

Year / Grade	2	3	4
1995	3.5%	0.0%	0.0%
1996	0.0%	5.8%	8.1%
1997	0.0%	5.7%	4.5%
1998	0.0%	1.2%	7.4%
1999	0.0%	0.0%	4.0%
2000	0.0%	0.0%	9.1%

Female

Year / Grade	2	3	4
1995	1.7%	0.0%	0.0%
1996	0.0%	1.2%	2.9%
1997	0.0%	3.4%	3.4%
1998	0.0%	1.7%	2.3%
1999	0.0%	1.1%	1.1%
2000	0.0%	0.0%	5.1%

Test Scores

Year / Grade	2	3	4
1995	0.0%	0.0%	0.0%
1996	0.0%	27.7%	0.0%
1997	0.0%	2.3%	26.1%
1998	0.0%	1.2%	26.7%
1999	0.0%	0.0%	6.3%
2000	0.0%	0.0%	1.1%

Table 7

Summary Statistics of Trimmed Samples

	Baseline Sample		Trimmed*		Trimmed (Free Lunch)		Trimmed (Non Free Lunch)	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
N	**34,450**		**16,811**		**7,459**		**9,352**	
SAT9 achievement outcomes								
Reading	42.12	*14.66*	43.83	*14.51*	40.70	*13.68*	46.35	*14.65*
Math	56.26	*19.01*	58.08	*18.95*	54.24	*17.84*	61.18	*19.27*
Lagged Reading	37.74	*15.49*	39.21	*15.64*	35.47	*14.70*	42.21	*15.71*
Lagged Math	56.99	*18.14*	58.46	*17.99*	54.94	*17.01*	61.30	*18.26*
Race, in percent								
White	17.67	*38.14*	21.83	*41.31*	12.40	*32.96*	29.47	*45.59*
Black	67.69	*46.77*	66.97	*47.03*	75.92	*42.76*	59.71	*49.05*
Hispanic	10.43	*30.57*	7.92	*27.01*	9.24	*28.96*	6.83	*25.23*
Asian	4.07	*19.76*	3.15	*17.47*	2.31	*15.01*	3.86	*19.26*
Other	0.14	*3.69*	0.13	*3.62*	0.13	*3.66*	0.13	*3.57*
Gender, in percent								
Male	48.86	*49.99*	46.76	*49.90*	44.85	*49.74*	48.22	*49.97*
Female	51.14	*49.99*	53.24	*49.90*	55.15	*49.74*	51.78	*49.97*
Academic Variables, in percent								
Attended Phila. Kindergarten	84.70	*36.00*	85.52	*35.19*	85.05	*35.66*	85.90	*34.81*
Free lunch eligible	53.32	*49.89*	49.47	*50.00*	100.00	*0.00*		
English language learner	3.18	*17.57*						
Special education	3.34	*18.08*						
Lagged behavior = D	10.92	*31.19*						
Lagged behavior = C	23.49	*42.39*	26.98	*44.39*	30.87	*46.20*	22.92	*42.04*
Lagged behavior = B	35.05	*47.71*	38.50	*48.66*	39.73	*48.94*	37.31	*48.37*
Lagged behavior = A	30.54	*46.06*	34.53	*47.55*	29.40	*45.56*	39.76	*48.94*
Student's census block								
Block percentage: white	30.00	*32.87*	32.90	*34.51*	25.31	*30.44*	36.24	*35.81*
Block percentage: poverty	14.14	*8.67*	13.38	*8.38*	15.52	*8.27*	12.12	*8.22*
Block percentage: house vacancy	12.74	*9.31*	12.28	*9.55*	14.47	*10.17*	11.00	*8.57*
Log of income (in dollars)	10.16	*0.45*	10.19	*0.45*	10.07	*0.44*	10.27	*0.43*
Teacher race, in percent								
White	80.58	*39.56*	78.47	*41.10*	78.76	*40.90*	78.29	*41.23*
Black	18.41	*38.75*	20.79	*40.58*	20.75	*40.56*	20.76	*40.56*
Hispanic	0.69	*8.25*	0.20	*4.49*	0.07	*2.59*	0.33	*5.74*
Asian	0.27	*5.19*	0.12	*3.45*	0.28	*5.30*	0.50	*7.06*
Other	0.06	*2.47*	0.42	*6.44*	0.13	*3.66*	0.12	*3.42*
Teacher gender, in percent								
Male	9.41	*29.19*	10.08	*30.11*	9.33	*29.09*	10.69	*30.90*
Female	90.59	*29.19*	89.92	*30.11*	90.67	*29.09*	89.31	*30.90*
Teacher skills								
Teacher experience (in years)	3.89	*7.73*	4.31	*8.16*	3.77	*7.51*	4.11	*8.20*
Teacher state certified (percent)	94.70	*22.40*	94.56	*22.68*	94.57	*22.66*	94.70	*22.00*
Teacher has a masters degree (percent)	14.33	*35.04*	16.30	*36.94*	13.82	*34.52*	13.76	*34.44*
Class size (head count)	28.79	*3.51*	28.91	*3.38*	28.63	*3.48*	29.13	*3.29*
Academic classroom characteristics								
Mean SAT9 reading score	36.35	*10.90*	37.60	*11.07*	35.45	*9.15*	40.27	*9.55*
Mean SAT9 math score	53.99	*13.84*	54.49	*12.06*	54.89	*8.82*	58.85	*9.71*
Other classroom characteristics (count)								
Free lunch	10.46	8.47	10.03	*8.25*	13.21	*8.13*	7.46	*7.44*
Behavior problems	1.80	1.74	1.58	*1.58*	1.83	*1.67*	1.37	*1.47*
English language learners	1.29	3.06	1.17	*2.74*	1.04	*2.59*	1.22	*2.73*
Special education	1.07	1.43	1.04	*1.31*	0.89	*1.25*	1.15	*1.35*
Female	10.65	3.61	10.47	*3.58*	10.45	*3.66*	10.52	*3.50*

*Note: The trimmed sample has been determined by non-special needs students who are also in equally distributed grades in a given year.

Table 8

Characteristics of Late Arrival Students and Their Non-Late Classmates

	Late Arrivals		Non-Late Classmates	
	Mean	***SD***	**Mean**	***SD***
Race, in percent				
White	17.52	*38.05*	17.14	*37.69*
Black	65.71	*47.51*	67.51	*46.83*
Hispanic	13.71	*34.43*	11.00	*31.29*
Asian	2.86	*16.68*	4.18	*20.02*
Other	0.19	*4.36*	0.16	*3.99*
Gender, in percent				
Male	53.33	*49.94*	48.83	*49.99*
Female	46.67	*49.94*	51.17	*49.99*
Acadmic Variables, in percent				
Free Lunch Eligible	44.58	*49.86*	52.77	*49.92*
English Language Learners	2.88	*16.75*	3.68	*18.82*
Special Education	3.62	*18.69*	3.90	*19.36*
Behavior Problem	10.10	*30.16*	8.52	*27.92*
SAT9, standardized score*				
Previous Year's SAT9 Reading	38.01	*17.46*	38.92	*15.51*
Previous Year's SAT9 Math	57.06	*19.39*	55.78	*18.91*

**Note: Test scores for late students are available in the data if a student was in the Philadelphia School District in the previous year and is late into a new school.*

Table 9a

Logistic Regression Results Predicting Probability of Classroom Assignment Based on Observable "Down-Branch" Characteristics in Illustration 1

		Chacteristics of Late Student				
		A	B	C	D	E
Classroom Characteristics	% Special Ed	3.221				
		(3.141)				
	% ELL	0.345	0.732			
		(1.989)	*(1.965)*			
	% Bad	-1.116	-0.262	-0.557		
		(1.819)	*(1.917)*	*(1.942)*		
	% FL	1.148	1.306	1.297	3.024	
		(1.859)	*(1.915)*	*(1.909)*	*(2.309)*	
	% Female	-1.844	-1.259	-1.137	-0.327	-2.651
		(2.246)	*(2.335)*	*(2.332)*	*(2.678)*	*(2.928)*
	Dummy: put in minimum classroom within grade	1.108 **	1.136 **	1.079 *	1.381 **	1.182 *
		(0.556)	*(0.559)*	*(0.565)*	*(0.657)*	*(0.687)*
	Controls for Teachers	Y	Y	Y	Y	Y
	Fixed Effects	Y	Y	Y	Y	Y
	N	329	313	309	244	204
	R-sq	0.17	0.17	0.17	0.20	0.18

**Notes: Robust standard errors are in italics*
Group A includes all late students
Group B is Group A minus special education late students
Group C is Group B minus ELL late students
Group D is Group C minus late behavior problems
Group E is Group D minus free lunch late students. Female remains the final discerning characteristic.

Table 9b

Logistic Regression Results Predicting Probability of Classroom Assignment Based on Observable "Down-Branch" Characteristics in Illustration 1, Including Previous Test Scores

Classroom Characteristics	Chacteristics of Late Student						
	A	B*	C*	D	E	F	G
% Special Ed	2.233						
	(4.160)						
% Low Scorers	0.037	-0.172					
	(1.076)	*(1.080)*					
% High Scorers	-0.617	-1.281	-1.305				
	(2.677)	*(2.737)*	*(2.204)*				
% ELL	4.486	4.346	3.599	2.353			
	(2.726)	*(2.682)*	*(2.507)*	*(3.395)*			
% Bad	-2.340	-2.149	-2.993	-0.259	-1.012		
	(2.396)	*(2.388)*	*(2.523)*	*(3.516)*	*(3.591)*		
% FL	0.919	1.147	1.720	4.057	4.268	1.177	
	(2.212)	*(2.226)*	*(2.323)*	*(3.189)*	*(3.146)*	*(4.252)*	
% Female	-1.823	-0.642	-0.049	-0.702	-0.197	2.010	-5.056
	(2.546)	*(2.586)*	*(2.657)*	*(4.005)*	*(3.913)*	*(4.482)*	*(6.246)*
Dummy: put in minimum classroom within grade	1.028 *	1.000 *	1.192 *	2.106 **	2.017 **	3.627 ***	2.969 **
	(0.589)	*(0.588)*	*(0.617)*	*(0.951)*	*(0.949)*	*(1.280)*	*(1.321)*
Controls for Teachers	Y	Y	Y	Y	Y	Y	Y
Fixed Effects	Y	Y	Y	Y	Y	Y	Y
n	238	230	217	128	126	105	81
R-sq	0.21	0.20	0.20	0.19	0.19	0.26	0.21

**Note: The results remain consistent when swapping the order of low scorer and high scorer. Results are also similar for low/high scorers in SAT9 math.*
Robust standard errors are in italics.
Group A includes all late students
Group B is Group A minus special education late students
Group C is Group B minus all low scoring late students
Group D is Group C minus all high scoring late students
Group E is Group D minus all ELL late students
Group F is Group E minus all behavior problem late students
Group G is Group F minus free lunch students. Female remains the final discerning characteristic.

Table 10

Baseline Model of Education Production

	Reading	Math		Reading	Math
Lagged test score	0.810 ***	0.499 ***	**Teacher male**	-0.145	-0.696
	(0.027)	*(0.034)*		*(0.493)*	*(0.506)*
Male	-0.437 ***	0.600 ***	**Teacher black**	-0.423	-1.540 *
	(0.118)	*(0.145)*		*(0.386)*	*(0.614)*
Black	-2.105 ***	-2.602 ***	**Teacher hispanic**	-0.352	1.192
	(0.266)	*(0.253)*		*(1.069)*	*(1.997)*
Hispanic	-1.215 ***	-1.636 ***	**Teacher asian**	0.820	6.650 *
	(0.326)	*(0.296)*		*(1.739)*	*(3.648)*
Asian	0.228	1.909 ***	**Teacher other**	-4.307 *	-9.816 *
	(0.395)	*(0.415)*		*(2.341)*	*(5.138)*
Other	-1.193	-1.270	**Teacher experience**	-0.118 *	-0.073
	(1.474)	*(1.604)*		*(0.063)*	*(0.084)*
Repeated current grade	5.621 ***	11.006 ***	**Teacher experience-sq**	0.005 *	0.003
	(0.428)	*(0.539)*		*(0.002)*	*(0.003)*
Young for grade	-0.101	0.079	**Teacher has masters**	0.442	0.671
	(0.175)	*(0.208)*		*(0.364)*	*(0.473)*
Repeat * young	-0.872	-1.923	**Teacher has certification**	-0.453	-0.795
	(0.484) *	*(3.683)*		*(0.556)*	*(0.628)*
Had K in Phila school district	-0.172	-0.087	**Class size**	0.335	0.274
	(0.184)	*(0.215)*		*(0.316)*	*(0.267)*
Special ed	-2.199 ***	-2.150 ***	**Class size - sq**	-0.004	0.001
	(0.339)	*(0.432)*		*(0.006)*	*(0.005)*
ELL	-0.907 *	-1.106 *	**Mean class lagged test score**	-0.449 ***	-0.453 ***
	(0.504)	*(0.593)*		*(0.081)*	*(0.102)*
Free lunch	-0.579 ***	-1.190 ***	**Mean class lagged test score - sq**	0.007 ***	0.002 **
	(0.143)	*(0.145)*		*(0.001)*	*(0.001)*
Last year behv: A	3.803 ***	4.929 ***	**Mean x individual lagged score**	-0.004 ***	0.003 ***
	(0.248)	*(0.254)*		*(0.001)*	*(0.001)*
Last year behv: B	1.626 ***	2.319 ***	**Class count of free lunch**	-0.121 ***	-0.155 **
	(0.202)	*(0.221)*		*(0.046)*	*(0.068)*
Last year behv: C	0.297	0.709 ***	**Class count of behv problems**	-0.276 ***	-0.263 **
	(0.218)	*(0.250)*		*(0.078)*	*(0.109)*
Census blc: % white	0.309	0.070	**Class count of ELL**	-0.050	-0.029
	(0.397)	*(0.373)*		*(0.070)*	*(0.092)*
Census blc: % pov	0.521	-0.807	**Class count of SE**	-0.146	-0.215 *
	(1.420)	*(1.484)*		*(0.097)*	*(0.113)*
Census blc: log(income)	-0.031	0.135	**Class count of females**	0.224 ***	0.166 ***
	(0.220)	*(0.276)*		*(0.048)*	*(0.059)*
Census blc: hh vac rate	-0.231	-1.524 *			
	(0.922)	*(0.902)*			
n	23,304	30,887			
R^2	0.61	0.63			

*Note: *** p<.01; **p<.05; *p<.1*

Robust standard errors adjusted for classroom clustering are in parentheses.

Regressions include school, year, and school-by-year fixed effects.

Table 11

Peer Effects for the Baseline and Trimmed Samples

	Reading		Math	
	Baseline	**Full Trimmed**	**Basline**	**Full Trimmed**
Mean class test score	-0.449 ***	-0.689 ***	-0.453 ***	-0.625 ***
	(0.081)	*(0.130)*	*(0.102)*	*(0.171)*
Mean class test score - sq	0.007 ***	0.010 ***	0.002 **	0.004 **
	(0.001)	*(0.002)*	*(0.001)*	*(0.002)*
Mean x individual lagged score	-0.004 ***	-0.005 ***	0.003 ***	0.002 **
	(0.001)	*(0.001)*	*(0.001)*	*(0.001)*
Class count of free lunch	-0.121 ***	-0.115 *	-0.155 **	-0.163 *
	(0.046)	*(0.064)*	*(0.068)*	*(0.098)*
Class count of behv problems	-0.276 ***	-0.240 ***	-0.263 **	-0.338 **
	(0.078)	*(0.122)*	*(0.109)*	*(0.183)*
Class count of ELL	-0.050	-0.072	-0.029	-0.097
	(0.070)	*(0.098)*	*(0.092)*	*(0.160)*
Class count of SE	-0.146	-0.209	-0.215 *	0.113
	(0.097)	*(0.136)*	*(0.113)*	*(0.172)*
Class count of females	0.224 ***	0.150 **	0.166 ***	0.174 **
	(0.048)	*(0.063)*	*(0.059)*	*(0.072)*
School, year, school-year fixed effects	Yes	Yes	Yes	Yes
n	23,304	10,732	30,887	14,751
R^2	0.61	0.62	0.63	0.65

*Note: *** p<.01; **p<.05; *p<.1; Robust standard errors adjusted for classroom clustering are in parentheses.*
Models also include all other covariates from Table 10, including school, year, and school-by-year fixed effects.

Table 12a

Estimating the Effect of Late Arrival Students on Non-Late Arrivals: Reading Achievement

	Strategy A	Strategy B	Strategy C
Late lagged test score	0.010 (b)	0.050	
	(0.117)	*(0.083)*	
Late lagged test score x own score	-0.004 (b)	-0.001	
	(0.003)	*(0.001)*	
Late student is free lunch	-0.142 ***	-0.143 *	-0.106 ***
	(0.016)	*(0.083)*	*(0.037)*
Late student is ELL	-0.280 ***	-0.744	-0.120
	(0.009)	*(1.077)*	*(0.129)*
Late student is special ed	-0.193 ***	0.268	0.054
	(0.005)	*(0.187)*	*(0.198)*
Late student is female	0.141 **	0.220 *	0.177 *
	(0.052)	*(0.121)*	*(0.091)*
Late student is behavior problem	-0.231 ***(b)	-0.223 *	
	(0.004)	*(0.111)*	
School, year, school-year fixed effects	School	Y	Y
n	421	8,514	10,713
R^2	0.63	0.63	0.30

*Note: *** p<.01; **p<.05; *p<.1; Robust standard errors adjusted for classroom clustering are in parentheses.*

(a) The coefficients for "late student is free lunch", "late student is ELL", "late student is special ed", and "late student is female" are all from seaparate regression models that control for lagged late student ability, lagged ability x own ability, indicator for late student being a behavior problem, mean class ability, mean class ability x own , and class counts of free lunch, behavior problems, and girls.

(b) These coefficients are from a "late student is free lunch" regression. However, there is a consistency for all four models run here.

Table 12b

Estimating the Effect of Late Arrival Students on Non-Late Arrivals: Math Achievement

	Strategy A	Strategy B	Strategy C
Late lagged test score	0.729 (b)	0.073	
	(0.952)	*(0.051)*	
Late lagged test score x own score	0.001 (b)	0.000	
	(0.003)	*(0.001)*	
Late student is free lunch	-0.293 ***	-0.199 **	-0.093 **
	(0.007)	*(0.009)*	*(0.004)*
Late student is ELL	-0.251 ***	-0.652	-0.030
	(0.007)	*(1.437)*	*(0.163)*
Late student is special ed	-0.174 ***	0.446	0.271
	(0.004)	*(0.276)*	*(0.175)*
Late student is female	0.263 ***	0.157 **	0.149 **
	(0.001)	*(0.061)*	*(0.077)*
Late student is behavior problem	-0.249 *** (b)	-0.219 *	
	(0.007)	*(0.120)*	
School, year, school-year fixed effects	School	Y	Y
n	574	8,317	14,751
R^2	0.70	0.66	0.35

*Note: *** p<.01; **p<.05; *p<.1; Robust standard errors adjusted for classroom clustering are in parentheses.*

(a) The coefficients for "late student is free lunch", "late student is ELL", "late student is special ed", and "late student is female" are all from seaparate regression models that control for lagged late student ability, lagged ability x own ability, indicator for late student being a behavior problem, mean class ability, mean class ability x own , and class counts of free lunch, behavior problems, and girls.

(b) These coefficients are from a "late student is free lunch" regression.
However, there is a consistency for all four models run here.

Table 13

Peer Effects for Trimmed Sample, Broken Out by Free Lunch and Non Free Lunch Students

	Reading		Math	
	Free Lunch	Non Free Lunch	Free Lunch	Non Free Lunch
Mean class test score	-0.611 ***	-0.697 ***	-0.735 ***	-0.315
	(0.172)	*(0.180)*	*(0.247)*	*(0.203)*
Mean class test score - sq	0.009 ***	0.011 ***	0.006 **	0.001
	(0.003)	*(0.003)*	*(0.002)*	*(0.002)*
Mean x individual lagged score	-0.007 ***	-0.006 ***	-0.001	0.002
	(0.002)	*(0.001)*	*(0.001)*	*(0.001)*
Class count of free lunch	-0.149 *	-0.110	-0.117	-0.223 **
	(0.089)	*(0.078)*	*(0.119)*	*(0.089)*
Class count of behv problems	-0.251 *	-0.233	-0.501 **	-0.225
	(0.139)	*(0.170)*	*(0.215)*	*(0.188)*
Class count of ELL	-0.130	-0.033	-0.198	-0.002
	(0.125)	*(0.117)*	*(0.195)*	*(0.195)*
Class count of SE	-0.139	-0.281 *	0.063	0.158
	(0.162)	*(0.143)*	*(0.215)*	*(0.196)*
Class count of females	0.125 *	0.169 **	0.146	0.185 **
	(0.073)	*(0.083)*	*(0.093)*	*(0.080)*
School, year, school-year fixed effects	Yes	Yes	Yes	Yes
n	5,505	5,227	7,244	7,470
R^2	0.58	0.66	0.60	0.69

*Note: *** p<.01; **p<.05; *p<.1; Robust standard errors adjusted for classroom clustering are in parentheses.*
Models also include all other covariates from Table 10.

Table 14

Same-Year Impact on Learning by a One Standard Deviation Increase in a Classroom Trait (Within-District)

	Reading			Math		
	Full Trimmed	Free Lunch	Non Free Lunch	Full Trimmed	Free Lunch	Non Free Lunch
Class count of free lunch ***(+1SD = 9 add'l)***	-0.405 *	-0.525 *	-0.387	-0.657 *	-0.472	-0.899 **
Class count of behv problems ***(+1SD = 2 add'l)***	-0.175 ***	-0.182 *	-0.169	-0.281 **	-0.416 **	-0.187
Class count of ELL ***(+1SD = 4 add'l)***	-0.092	-0.166	-0.042	-0.141	-0.289	-0.003
Class count of SE ***(+1SD = 2 add'l)***	-0.125	-0.083	-0.168 *	0.077	0.043	0.108
Class count of girls ***(+1SD = 4 add'l)***	0.226 **	0.188 *	0.254 **	0.299 **	0.251	0.318 **

*Note: *** p<.01; **p<.05; *p<.1*

Table 15

Initial Impact: The Effect on Learning Months over Three Years (Within-District)

	Reading			Math		
	Full Trimmed	Free Lunch	Non Free Lunch	Full Trimmed	Free Lunch	Non Free Lunch
Class count of free lunch ***(+1SD = 9 add'l)***	-1.125 *	-1.496 *	-1.118	-1.345 *	-1.126	-1.909 **
Class count of behv problems ***(+1SD = 2 add'l)***	-0.523 ***	-0.559 *	-0.527	-0.621 **	-1.072 **	-0.428
Class count of ELL ***(+1SD = 4 add'l)***	-0.315	-0.580	-0.149	-0.355	-0.849	-0.007
Class count of SE ***(+1SD = 2 add'l)***	-0.455	-0.311	-0.637 *	0.206	0.135	0.301
Class count of females ***(+1SD = 4 add'l)***	0.653 **	0.556 *	0.762 **	0.637 **	0.623	0.703 **

*Note: *** p<.01; **p<.05; *p<.1*

Table 16

Continual Impact: The Effect on Learning Months over Three Years (Within-District)

	Reading			Math		
	Full Trimmed	Free Lunch	Non Free Lunch	Full Trimmed	Free Lunch	Non Free Lunch
Class count of free lunch ***(+1SD = 9 add'l)***	-2.360 *	-3.110 *	-2.312	-3.499 *	-2.768	-4.900 **
Class count of behv problems ***(+1SD = 2 add'l)***	-1.254 ***	-1.329 *	-1.247	-1.765 **	-2.658 **	-1.203
Class count of ELL ***(+1SD = 4 add'l)***	-0.754	-1.377	-0.353	-1.0104	-2.104	-0.019
Class count of SE ***(+1SD = 2 add'l)***	-1.091	-0.738	-1.505 *	0.587	0.335	0.846
Class count of females ***(+1SD = 4 add'l)***	1.565 **	1.321 *	1.803 **	1.811 **	1.544	1.976 **

*Note: *** p<.01; **p<.05; *p<.1*

Table 17

Same-Year Impact on Learning of a One Standard Deviation Increase in a Classroom Trait (Within-School)

	Reading			Math		
	Full Trimmed	**Free Lunch**	**Non Free Lunch**	**Full Trimmed**	**Free Lunch**	**Non Free Lunch**
25th percentile school (FL) ***(+1SD = 4 add'l)***	-0.191 *	-0.248 *	-0.183	-0.310 *	-0.223	-0.425 **
50th percentile school (FL) ***(+1SD = 9 add'l)***	-0.405 *	-0.525 *	-0.387	-0.657 *	-0.472	-0.899 **
75th percentile school (FL) ***(+1SD = 7 add'l)***	-0.335 *	-0.434 *	-0.320	-0.543 *	-0.390	-0.743 **

*Note: *** p<.01; **p<.05; *p<.1*

Table 18

Initial and Continual Impacts' Effects on Learning Months (Within-School)

	Initial Impact					
	Reading			Math		
	Full Trimmed	Free Lunch	Non Free Lunch	Full Trimmed	Free Lunch	Non Free Lunch
Schools at 25th percentile ***(+1SD = 4 add'l)***	-0.500 *	-0.665 *	-0.497	-0.598 *	-0.500	-0.849 **
Schools at 50th percentile ***(+1SD = 7 add'l)***	-1.125 *	-1.496 *	-1.118	-1.345 *	-1.126	-1.909 **
Schools at 75th percentile ***(+1SD = 7 add'l)***	-0.875 *	-1.164 *	-0.869	-1.046 *	-0.876	-1.485 **

	Continual Impact					
	Reading			Math		
	Full Trimmed	Free Lunch	Non Free Lunch	Full Trimmed	Free Lunch	Non Free Lunch
Schools at 25th percentile ***(+1SD = 4 add'l)***	-1.049 *	-1.382 *	-1.028	-1.400 *	-1.107	-1.960 **
Schools at 50th percentile ***(+1SD = 7 add'l)***	-2.360 *	-3.110 *	-2.312	-3.499 *	-2.768	-4.900 **
Schools at 75th percentile ***(+1SD = 7 add'l)***	-2.097 *	-2.764 *	-2.055	-2.974 *	-2.172	-4.172 **

*Note: *** p<.01; **p<.05; *p<.1*

Table 19

The Effects of Peers on Non-Academic Outcomes

	Dependent variable: indicator of good behavior in year *t*			Dependent variable: rate of unexcused absences in year *t*		
	Baseline	Trimmed	Late Arrival [(a)]	Baseline	Trimmed	Late Arrival [(a)]
Lagged outcome measure [(b)]	-0.763 ***	-0.869 ***	-0.768 ***	0.140 ***	0.115 ***	0.140 ***
	(0.043)	*(0.034)*	*(0.043)*	*(0.006)*	*(0.009)*	*(0.006)*
Mean class test score	0.030 ***	0.026 *	0.029 ***	-0.002	-0.003	-0.003
	(0.011)	*(0.015)*	*(0.011)*	*(0.002)*	*(0.003)*	*(0.002)*
Mean class test score - sq	-0.001 ***	0.000 *	-0.001 ***	0.000	0.000	0.000 **
	(0.000)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*
Mean x individual lagged score	0.000 **	0.000	0.000 **	0.000	0.000 ***	0.000 ***
	(0.000)	*(0.000)*	*(0.000)*	*0.000*	*(0.000)*	*(0.000)*
Class count of free lunch	0.010	0.001	0.009	0.002 ***	0.002 **	0.002 ***
	(0.006)	*(0.009)*	*(0.006)*	*(0.001)*	*(0.001)*	*(0.001)*
Class count of behv problems	-0.044 ***	-0.055 ***	-0.044 ***	-0.001	-0.002	-0.002
	(0.013)	*(0.017)*	*(0.013)*	*(0.001)*	*(0.002)*	*(0.001)*
Class count of ELL	-0.043 ***	-0.054 ***	-0.041 ***	-0.004 ***	-0.006 **	-0.005 ***
	(0.012)	*(0.016)*	*(0.012)*	*(0.001)*	*(0.003)*	*(0.001)*
Class count of SE	-0.027 *	-0.019	-0.027 *	-0.001	-0.003	-0.001
	(0.015)	*(0.020)*	*(0.015)*	*(0.002)*	*(0.003)*	*(0.002)*
Class count of females	0.014 *	0.017 **	0.012	-0.001	0.000	0.000
	(0.008)	*(0.009)*	*(0.008)*	*(0.001)*	*(0.001)*	*(0.001)*
School, year, school-year fixed effects	Y	Y	Y	Y	Y	Y
n	49,201	30,880	49,103	45,013	23,715	44,923
R^2	0.24	0.20	0.24	0.38	0.42	0.39

*Note: *** p<.01; **p<.05; *p<.1; Robust standard errors adjusted for classroom clustering are in parentheses.*
Models include all covariates from each analogous specification for reading and math achievement.
(a) Late student regressions, "other peer effects" are constructed as the classroom having this certain type of late student rather than total class counts.
(b) Lagged outcomes for behavioral regressions are three-fold: A, B, or C grade in previous year. For simplicity of presentation, only lagged grade of C is presented here.

CHAPTER 2

PROGRAMS, PEOPLE, AND PROPERTY: EXAMINING THE INSTITUTIONAL-LEVEL FACTORS OF URBAN SCHOOL EFFECTIVENESS

Chapter Abstract

The large number of school-level effects evaluated in the literature yields evidence of important consequences for how institutional characteristics can influence school effectiveness. Given the diverse number of studies on unique school effects, this paper attempts to consolidate this recent evidence on the consequences of school resources on school quality. Specifically, this study implements the theory of educational production from the economics of education literature and extends the model to an empirical understanding of which school-level resources relate to school quality, holding constant student, teacher, classroom, and neighborhood information. By conducting analysis on a school-level dataset of 175 elementary schools within the School District of Philadelphia over the years 1997 through 2000, this study provides evidence that a range of school-level resources – as broken out by programs, people, and property – have significant relationships to school quality in both SAT9 reading and math test subject areas.

Introduction

In the United States, urban schools are characterized by low levels of educational attainment, high dropout rates, and graduates who are inadequately prepared for postsecondary opportunities (Tighe, Wang, & Foley, 2002). With the number of children in poverty rising coupled with evidence that high-poverty schools are disproportionally composed of ethnic and racial minorities,[21] urban districts are increasingly being populated with students at the lowest levels of academic achievement. As such, researchers and policymakers have identified urban minority children as particularly vulnerable to educational failure, and over the last few decades, evidence indicates that America's city schools need serious improvement.

Simultaneously over the last several decades, the traditional notion of improving urban schooling performance – that the way to improve student achievement is through an increased allocation of funding – has been challenged by researchers. While student expenditures have risen dramatically, it is not clear whether or not achievement has risen to match. Several studies have specifically examined the effect of school financial resources on schooling quality and student achievement (e.g. Ehrenberg & Brewer, 1994, 1995; Hanushek, 1986, 1996). However, they found that improving school resources, such as increased per pupil spending, did not necessarily increase academic performance on standardized exams.

[21] As an example, in the School District of Philadelphia, approximately 65% of the student population is Black.

Nonetheless, schools continue to be attributed with influencing student outcomes (Firestone, 1991; Mortimore et al., 1988; Reynolds & Creemers, 1990; Rutter et al., 1979). To be specific, Mortimore (1991) has defined an effective school as one in which students progress further than might be expected from consideration of its student population. That is, an effective school adds extra value to its students' outcomes in comparison with other schools serving similar students. By contrast, an ineffective school has students who progress less than expected.

Provided that some schools can be credited with being capable of impacting student outcomes more efficiently than others, and yet increased financial resources do not seem to provide a definitive answer as to how, both academic researchers and policy makers should be been asking: what will make a difference? Examining this question is the scope of this paper. In particular, this study utilizes school-level variables to assess the average quality of individual school performance over the years 1997 to 2000 for public elementary schools in the School District of Philadelphia. This paper asserts that, holding constant student, teacher, classroom, and neighborhood characteristics, there still remain school-level resources that can increase a school's effectiveness. Three overall categories of school-level inputs are proposed in this paper: school-wide programs, personnel resources, and school environment.

Using an empirical model of education production, this study constructs a quantifiable measure of school quality based on the results from a related set of multilevel analyses in Gottfried and Inman (2010). From this starting-point, this study evaluates the effects of institutional-level resources on two measures of school quality derived from student-level standardized testing performance (SAT9) in reading and math subjects. The results indicate that institutional-level resources are significantly related to school quality across three categories (programs, personnel, and school environment) and within both testing subject areas. While there is some consistency across both reading and math school effectiveness, the results also indicate that differentiating between subject tests is crucial: school resources may provide distinctive institutional effects depending on the testing area itself.

Background

Some studies have shown little evidence between academic performance and school inputs (Betts, 1995; Grogger, 1995; Hanushek, 1986). However, more recently, an increasing body of literature in both education and economics of education fields has found institutional-level variables to exert significant effects on the production of educational outcomes. The studies relevant to this paper involve those analyses of educational outcomes as they relate to school-level inputs, such as programs, personnel, and school environment as mentioned above.

First, this paper evaluates how school quality is related to having school-wide programs. As an example of this analysis of this relationship, Rolstad, Mahoney, and Glass (2005) conducted a meta-analysis of program effectiveness on research on English language learners (ELL). Their results indicated that school-wide bilingual education programs are effective in promoting achievement and that policy should encourage schools without these resources to develop and implement ELL programs. As another example, Hanushek, Kain, and Rivkin (2002) employed data from the UTD Texas Schools Project to track special education students who transferred in and out of targeted programs, thereby providing a measure of program effectiveness over time for the same student. They found that schools with special education programs boosted effectiveness in mathematics achievement for special education students without detracting from non-special needs students.

As an example of the effects of non-special needs school-level programming on academic effectiveness, the education literature has suggested that exposure to music in elementaryschool may improve current and future educational outcomes. Moreover, much of this literature correlates music exposure and mathematics success. For instance, Gardiner, Fox, Knowles, and Jeffrey (1996) found that those first and second grade students who received seven months of supplementary music classes at school achieved higher standardized math scores than children in the control group who did not receive the treatment. Similarly, Granziano, Peterson, and Shaw (1999) found that the mathematical reasoning scores of children who received music instruction were significantly higher than their counterparts.

A second set of school-level inputs includes personnel, and thus a second set of studies on school-level inputs relates human resources to school quality. Several papers have demonstrated that shared administrative responsibility among the principal and other administrative staff in the school can lead to increased school effectiveness (Sammons et al., 1995). That is, by expanding the responsibilities to a cabinet of administrative personnel rather than having to rely on the time-constraints of a single principal, there may be increased efficiency in school operations. Similarly, Flessa (2003) reported that having a specialized staff in a school's governance structure, such as a designated community liaison or disciplinarian, allows the principal to focus on envisioning and executing school curriculum and student learning. In the same vein, Grubb and Flessa (2006) reported that an expanded management staff may lead to closer attention being paid to instructional practices, for which principals complained that they often do not have time. If the principal can free up his or her time for instructional practices, this means that another manager, such as an assistant principal, can pay attention to support and disciplinary services, neglected in many schools.

Other studies have examined the effects of school-level personnel within the context of health and educational outcomes. For instance, Allen (2003) examined the relationship between health-related student issues and test performance for schools which had nurses versus those that did not. The results indicated that elementary schools with nurses had fewer absent children for medical reasons than did schools without nurses. The implications of his findings were that declines in medical leaves increased in-classroom instruction, which the author suggested led to a decrease in squandered schooling resources from otherwise absent students. Similarly, Guttu, Engelke, and Swanson (2004) evaluated the number of school nurses in public schools in 21 counties within North Carolina and determined that the presence of a school nurse increased medical screenings and follow-ups for student with health issues. The results indicated a decline in the spread of sickness within schools as well as an increased capability of nursing personnel to match sick students with particular educational needs.

Finally, some literature has focused on those school-level variables relating to environment. For instance, Tighe, Wang, and Foley (2002) used multilevel models and found that total student enrollment was related to a higher degree of aggregate school obstacles to learning. Offenberg (2001) found a relationship between school structure and academic attainment. Specifically, he relied on a series of natural experiments in the School District of Philadelphia to determine that on average, students in K-8 schools had higher levels of achievement than students in middle schools. Byrnes and Ruby (2004) used multilevel modeling to examine the educational outcomes of five cohorts of students in Philadelphia. Consistent with other literature, the authors also found a higher level of achievement for students in K-8 schools, compared to their counterparts in middle schools.

The diverse number of institutional effects studied in the literature yield evidence of important consequences of the institutional arrangements on school effectiveness. Therefore, given the particularly diverse number of separate school effects studied in the field, this paper attempts to consolidate this recent evidence on the institutional predictors of school quality. In particular, this study unifies these different school effects into a single model of educational production. The model does so while simultaneously holding constant all student, teacher, classroom, and neighborhood characteristics. The results are pertinent because school quality is actualized as a measure of standardized test achievement, per school, for both reading and math.

Method

To examine the effects of institutional-level resources, this study begins with the standard education production function, as first developed by Summers and Wolfe (1977), Henderson, Mieszkowski, and Sauvageau (1978) and later revised by Todd and Wolpin (2003). This model evaluates the relationship between school inputs and various output measures of achievement. In this regard, academic outcomes are comparable to learning as a technology.

A basic form of the model is expressed as follows:

$$A_{it} = f(G_{it}, F_{it}, N_{it}, T_{it}, C_{it}, S_{it}). \qquad (1)$$

where A represents student achievement; G, the student's own characteristics; F, family characteristics; N, neighborhood characteristics; T, teacher characteristics; C, classroom characteristics; and S, school characteristics. The subscripts indicate that it is possible to add time components to the inputs and outputs of equation. Doing so implies a contemporaneous model of student achievement: a student's test score in year *t* is a function of the influences of all input vectors in year *t*.

However, rather than assuming that a current year's achievement outcome is strictly a function of current inputs, it is possible to enrich the education production function model to include inputs from previous time periods. In fact, it is theoretically possible to include all time periods for which the student is in school. This model is known as the historical model of education production. To derive this full historical, cumulative-learning model, it is important to make an initial assumption, as developed by Todd and Wolpin (2003), that achievement in the initial period of schooling is a function of the student's natural endowment and family inputs provided prior to the period in which the student enters his or her first year of schooling. Those family inputs in the previous period of initial schooling are described as follows:

$$F_0 = f_0(G_i), \tag{2}$$

where F_0 is family inputs in before-schooling period 0 and G_i is student *i*'s natural endowment. Because the student has not yet enrolled in school in period 0, there is no academic achievement information for the student, and hence the family at this point can only adjust its inputs to the student's learning process based on their direct observations of the student's ability level, G.

Then, in the first period of schooling, student achievement is a function of ability G, family inputs F, and contemporaneous school inputs S:

$$A_1 = f_1(G, F_0(G), F_1(G), S_1) \tag{3}$$

Note that in this first year of school learning, school inputs do not adjust to the child's ability. In practice, this is demonstrated by the fact that students are more-often-than-not randomly assigned to a classroom in the starting grade that the school offers, either kindergarten or first grade.

In subsequent periods, however, schools and parents can potentially adjust their respective inputs, based on the student's reported achievement performance from the previous period. This is demonstrated by the fact that family and school inputs in year 2, respectively F_2 and S_2, are functions of A_1, the previous year's achievement:

$$A_2 = f_2(G, F_0(G), F_1(G), F_2(G, A_1), S_1, S_2(A_1)) \tag{4}$$

Iterating this process for each year of schooling provides the following education production function for a student in a given year of schooling *t*, which includes both contemporaneous and historical information:

$$A_t = f_t(G, F_0(G), F_1(.) \ldots F_t(.), S_1, S_2(.) \ldots S_t(.)) \tag{5}$$

This model states that achievement, for a student in a year *t*, is a function of a student's ability level (which does not change over time), the family's inputs in the year prior to schooling and through year *t*, as well as school inputs from the first year of schooling through year *t*.

The linear representation of the education production function in equation (5) hypothetically requires all current and historical inputs pertaining to a student's schooling history. However, it is a difficult and challenging task to acquire all to estimate a fully specified historical education production function. As such, the widespread solution to this problem is to take the first difference of equation (5). The result is known as the value added specification, where all input requirements reduce to current inputs plus achievement from the *t-1* period:

$$a_{ijkt} = \beta_0 + \beta_1 a_{ijk(t-1)} + \beta_2 G_{it} + \beta_3 F_{it} + \beta_4 N_{jkt} + \beta_5 T_{jtk} + \beta_6 C_{jt} + \beta_7 S_{kt} + \varepsilon_{ijtk} \tag{6}$$

where achievement a is for student *i* in classroom *j* in school *k* in year *t* as the dependent variable and in year *t-1* as a lagged measure of ability[22]; G is a vector of student-level characteristics in year *t*; F is a function of family inputs for student *i* in year *t*; N includes neighborhood characteristics for student *i* in year *t*; T are teacher effects in classroom *j* in school *k* in year *t*; C are classroom-specific characteristics for classroom *j* in year *t*; S are school characteristics and institutional-level resources in year *t*; and the error term ε includes all unobserved determinants of achievement.

According to Goldhaber and Brewer (1996), school level resources S_{kt} may consist of school-, teacher- and classroom-specific variables that relate to institutional effectiveness. Further, Mundlak (1961) asserted that a linear term in the production function that pertains to institutional-level inputs must also contain the effects of management. If equation (5) is correctly specified with these institutional-level resources, then ordinary least squares yields consistent estimates of β_1 through β_7.

However, suppose the vector of schooling resources S_{kt} can be composed into two parts: observable characteristics, Z_1, such as school size, and unobservable characteristics, Z_2, such as managerial effort. Further, suppose that Z_1 can be included in the model whereas Z_2 cannot be, possibly due to the lack of measure of managerial or other school-level influences (Mundlak, 1961). Hence, the true model is:

[22] Although some of the literature implements the difference between current and lagged achievement as the dependent variable, this study places lagged achievement on the right hand side of the equation in order to avoid restricting the parameter to a value of one (Wolpin & Todd, 2003).

$$a_{ijkt} = \beta_0 + \beta_1 a_{ijk(t-1)} + \beta_2 G_{it} + \beta_3 F_{it} + \beta_4 N_{jkt} + \beta_5 T_{jtk} + \beta_6 C_{jt} + \gamma Z_{1kt} + \delta Z_{2kt} + \varepsilon_{ijtk} \quad (7)$$

However, because of unobservable information, the model estimated is:

$$a_{ijkt} = \beta_0 + \beta_1 a_{ijk(t-1)} + \beta_2 G_{it} + \beta_3 F_{it} + \beta_4 N_{jkt} + \beta_5 T_{jtk} + \beta_6 C_{jt} + \gamma Z_1^* + \nu_{ijtk} \quad (8)$$

where the error term now consists of unobservable institutional-level characteristics as well as a random error component.

The omission of Z_2 can cause two potential problems. First, the total effect of schooling resources may be understated because omitted factors may not be included in the explained portion of the variance of student achievement. Second, the equation may yield biased estimates of the effects of particular schooling resources on student outcomes. If repeated measures over time are available, the standard technique to account for this omitted variable bias is to estimate a fixed effects model (Goldhaber & Brewer, 1996). In a sample of I = 1...N schools with T observations per school, it is possible to estimate the following equation:

$$a_{ijkt} = \beta_0 + \beta_1 a_{ijk(t-1)} + \beta_2 G_{it} + \beta_3 F_{it} + \beta_4 N_{jkt} + \beta_5 T_{jtk} + \beta_6 C_{jt} + \gamma_{ijkt}. \quad (9)$$

The error term is decomposed as

$$\gamma_{ijkt} = \delta_j + \omega_t + \upsilon_{kt} + \varepsilon_{ijkt} \tag{10}$$

where δ_j are school fixed effects, ω_t are year fixed effects, υ_{kt} are school-by-year fixed effects, and (ε_{ijkt}) is a random error capturing individual variations over time as well as a class-specific random component that is common to all members of the same classroom. Empirically, this latter component of the error structure is estimated as robust standard errors adjusted for classroom clustering.

In more detail, school fixed effects δ_j control for the influences of school resources by capturing systematic differences across each unique institution. By, in essence, holding constant those time invariant school-specific characteristics, such as curriculum, school neighborhood, leadership, organization, and hiring practices, the school fixed effects control for school-level effectiveness, or quality. In his estimation of institutional-level fixed effects using the production function, Mundlak (1961) assumed that management (among other related institutional resources) did not change over time during the period of estimation. In other words, he restricted his analyses to the use of school (and year) fixed effects.

However, with the implementation of repeated measures in this study, incorporating school-by-year fixed effects – in addition to school and year fixed effects – avoids having to rely on Mundlak's (1961) assumption of strict time-invariance of unobservable institutional-level characteristics. Because school-by-year fixed effects allows for the model to control for systematic year-to-year changes at the school-level, equation (10) accounts for time invariance in school factors, such as those related to leadership and curriculum changes. In other words, any pattern of school-level effectiveness that is unique to a particular institution in a given year will be estimated (and therefore held constant) in addition to those time-invariant factors that contribute to a school's quality.

While this fixed effects framework does quantify the total effects of time invariant and time variant school-level resources into two tangible components, this specification does not allow for the identification or differentiation in the details of the effects of particular observable resources in vector Z_1. As such, the school-level inputs to the education production function that are particularly of interest for policy purposes cannot be ascertained because the inclusion of Z_1 along with school and school-by-year fixed effects would lead to perfect collinearity. Consequently, for the purposes of this paper, it is necessary to take an additional step in order to discern between the influences of specific school-level resources.

The solution is two-part. The first step is to estimate equation (9) including school, year, and school-by-year fixed effects (equation 10) in order to obtain estimates of the total effects of institutional effectiveness, or quality. Gottfried and Inman (2010) have executed this first step in a related study on the estimation of classroom peer effects. The authors implemented quasi-experimental methods on elementary school students to obtain estimates on all covariates in equations (9) and (10). In addition to providing student, classroom, teacher, and neighborhood estimates under this empirical framework, the authors derived estimates of total school effectiveness (i.e., coefficients for school and school-by-year fixed effects). As such, it is possible to estimate a second relationship – one between school effectiveness and school-level inputs.

The second step in this analysis involves implementing the fixed effects estimates from step one on a secondary regression. In this model, the estimates of school and school-by-year fixed effects are combined into a single measure of school quality and are subsequently regressed on observable school-level variables Z_1 (Goldhaber & Brewer, 1996; Rausch 1993). Since the school and school-by-year fixed effects are derived from the estimation of equations (9) and (10), the following expression presents the model to be explored in this study:

$$\delta_k + \upsilon_{kt} = Q_{kt} = f(Z_1) + e_{kt} \tag{11}$$

Empirically derived from stage one, the dependent variables of this second regression account for unobserved characteristics of a particular school environment, holding constant student variables, neighborhood information, classroom environments, and teacher variables. After controlling for these covariates, what remained to be estimated in stage one was an error structure comprised of school and school-by-year fixed effects in which the coefficient estimates "added" or "subtracted" values to student academic outcomes. These fixed effects, in essence, have provided measures of institutional-level effectiveness in a given year and over time independent of any student in the school. In other words, the dependent variable provides a quantifiable measure of school quality.

The task of evaluating school quality is conducted with an analogous education production function specification that now relates the output of education at the school-level to various school-level inputs. Like the student-level education production function explaining student educational outcomes through a series of inputs, the school-level education production function also has its roots in the economics of education literature (Cohn, 1968; Hanushek, 1986; Lee & Barro, 1997; Riew, 1966). In this study, it is expressed as follows:

$$Q_{kt} = Q(P_{kt}, R_{kt}, E_{kt}) + e_{kt} \tag{12}$$

where Q denotes school quality (which is the measure of school effectiveness in this study), based on the school and school-by-year fixed effects estimates above. As an output, school quality is derived from a multitude of institutional-level inputs, and include: P, which are school-wide programming resources; R, as personnel and governance resources; E, describing the school environment; and error term *e* incorporating unmeasured factors affecting school quality that are assumed to be independent and identically distributed.

The theoretical input-output process of equation (12) is represented empirically as the linear specification in equation (13):

$$Q_{kt} = \beta_0 + \beta_1 P_{kt} + \beta_2 R_{kt} + \beta_3 E_{kt} + e_{kt} \qquad (13)$$

where Q_{kt} is the sum of the school and school-by-year fixed for school k in year t.

Data

The analysis of school quality is facilitated by an unusually unique and comprehensive dataset of school-level characteristics. The data encompass elementary schools in the School District of Philadelphia over the academic years spanning 1997 through 2000. In sum, the sample contains 174 schools with elementary grades, either K-5 or K-8. Over the time span of the data, there are approximately 675 school-year observations to be implemented for the analysis of reading or math school effectiveness. Information regarding school characteristics was provided by the administrative offices of the School District of Philadelphia.

Dependent Variables

The set of dependent variables are constructed as the sum of school and school-by-year fixed effects for SAT9 reading and math based on the quasi-experimental regression results of student achievement from Gottfried and Inman (2010). These measures were derived from analyses of a student- and classroom-level dataset linked to the current school dataset by de-identified institutional and year information. This student-level dataset contains student achievement measures and vectors of student, teacher, classroom, and neighborhood characteristics as well as coded identification for school, classroom, grade, and academic year. This data were comprised of all students within the entire elementary school system within the School District of Philadelphia. This dataset in its entirety consisted of a total of N = 97,007 student observations within elementary grades over the time period 1994/1995 through 2000/2001. Thus, the coefficients on these fixed effects serve as the link between the school-level dataset and previous work analyzing student-level achievement in the School District of Philadelphia. Appendix D provides information on the student-level dataset from Gottfried and Inman (2010).

Proceeding forward, all analyses rely on the fact that student, teacher, classroom, and neighborhood characteristics are held constant by the mere nature of the construction of the measure of school quality from the Gottfried and Inman (2010) analyses. Table 1 presents descriptive statistics of dependent and independent variables used in this study. The dependent variables are measures of school quality, as defined in each testing subject as the sum of school and school-by-year fixed effects. These measures provide indicators of year-specific effectiveness in achievement, based on performance on the SAT9 reading and math of school k in year t.[23] The measures of school quality across both testing subject areas have a mean of 0. Hence, the average school in the district has zero school

[23] The unit of measure of SAT9 used in this paper is the Normal Equivalent Curve (NCE). Appendix D provides the mean SAT9 NCE for reading and math for elementary school students in the School District of Philadelphia.

quality, in this analysis. Less-than-average effective schools will have negative school quality, whereas greater-than-average effective schools have a positive measurement of school quality.

Note that the correlation between reading school quality and math school quality is approximately 0.40. This implies that a school with a strong reading quality tends to also have strong quality in mathematics. The same can be stated about lower performing schools: those at the lower end of the performance spectrum tend to perform worse across both testing subjects.

Independent Variables

As laid-out in the econometric strategy, institutional-level variables relating to school effectiveness fall into one of three categories in this study: programs, personnel resources, or school environment. First, there are several variables related to school-wide programming. As presented in Table 1, these programs include music, language skills,[24] and English instruction for non-native speakers ("ELL"). Each program variable is a binary indicator as to whether or not a school has a designated program in music, language, or ELL, respectively.

Table 2 first presents the results of three logistic regressions related to school-wide programming. The dependent variables are binary indicators for whether or not a school

[24] A language skills program refers to students who require additional language needs in the English language (e.g., speech therapy).

has a music, language skills, or ELL program, respectively. The independent variables include institutional-level measures of student demographics and special needs. Following the methodology of Sacerdote (2000), the binary indicators for school programs are regressed on school characteristics in order to determine if a significant predictive relationship is present. If there is, then this may be evidence of non-random assignment of programs to schools. However, the results are methodologically consistent with Sacerdote (2000): the lack of significant coefficients on school characteristics in Table 2 indicates that no systematic relationship exists between school demographic characteristics and school-wide programs.

A second set of independent variables includes personnel resources. First is the number of special education teachers per special education student in a given school. This variable describes the breadth of special education resources in a school: a larger value signals more available school-wide resources for special education students. On average, an elementary school in Philadelphia has approximately 3 special education teachers.

A second set of personnel variables relate to disciplinary resources and include indicators for whether schools have assistant principals or safety officers. However, to ascertain a measure of the breadth of disciplinary resources, the total number of school disciplinarians (constructed as the sum of assistant principals and safety officers) is divided by the number behavior problems per school[25]. This measure indicates how much of a disciplinary resource can be allotted per behavior problem. A final group of personnel resources pertain to additional staff pertaining to parental support and outreach and include indicators as to whether a school, in a given year, had a school nurse or school community liaison.

As with school program indicators, Table 2 also presents regression results pertaining to the relationship between school personnel resources and student body demographics. To avoid collinearity with student characteristics, the institutional-level dependent variables are indicators as to whether a school had a particular staff member (as opposed to ratios of staff members to students), so that student characteristics could serve as predictors. Like with programs, the results indicate a lack of significance between school characteristics and human resources. There does not appear to be evidence, thus, of a systematic matching relationship between school-level academic or demographic indicators and institutional personnel.

Table 1 finally presents the means and standard deviations for variables measuring school environment. The first is a measure of the number of teachers per student per school. This metric provides an indicator in the breadth of adult school environment. On average, there are approximately 590 students per school and 20 teachers. Second, a

[25] Students are deemed behavior problems in school-year *t* if they received a grade of D in behavior on their report cards from year *t-1*.

measure of the school's capital is constructed as the physical square footage per student. Third, a binary variable indicates if a school is K-8 (versus K-5).

Results

In this section, school quality, as represented by a measure of school and school-by-year fixed effects, is regressed on the three categories of school-level independent variables: programs, personnel resources, and school environment. This empirical specification allows for the evaluation of the effect of each variable on school quality, holding constant the effect of students, teachers, classrooms, and neighborhoods. The analysis is conducted twice, once for school quality in reading and once for math. Doing so enables for differentiation of school effects based on two subject areas.

School Quality in Reading

Table 3 presents parameter estimates, robust standard errors, and approximate *p* values from fitting the model in Equation (13) for school quality in SAT9 reading. For comparability, the table presents two versions of the results. The first column of estimates provides unstandardized regression coefficient estimates, in which the results correspond to absolute point gains or losses in the school quality for school *k* in year *t*. The second series of estimates presents the standardized regression coefficients for reading, thereby allowing for the evaluation of effect sizes. Standardized betas represent the magnitude of the unique effect of a particular independent variable on the dependent variable, controlling for the effects of other independent variables in the model. Because the variables for the analysis in this column are standardized, it is possible to compare the relationship of the direct effects of each independent variable with each other.

Of the variables pertaining to school-level programming, one program is significant in its relationship to school quality. Specifically, schools with a language skills programs have higher school quality in reading (β = 2.10, $p<0.10$) than do schools without such programs, holding all else equal. The standardized estimates suggest an effect size of 0.09σ.

The results of personnel variables suggest several significant relationships between human resources and school quality in reading. To begin, schools with more disciplinary resources per behavior problem have higher levels of school quality in reading than do schools with fewer disciplinary resources per behavior problem (β = 24.28, $p<0.05$). Recall that this variable is constructed as number of disciplinarians per behavior problem per school, and the mean from Table 1 suggests that about 1-percent of a school's disciplinary resources is allotted per behavior problem. Thus, the large coefficient suggests that if a school increased its disciplinary resources per student by 10-percent from 1-percent to 11-percent, this would be associated with an approximate 2.5-point increase in school quality in reading. The standardized coefficient of this variable suggests that a one standard deviation increase in school disciplinary resources per behavior problem is related to a 0.06 standard deviation increase in school quality in reading.

In conjunction with the result on disciplinary resources, the main effect of having an assistant principal is also statistically significant. The results indicate that schools with assistant principals have higher quality in reading than do schools without assistant principals (β = 3.41, $p<0.10$).[26] This corresponds to an effect size of approximately 0.11σ. Because the effect of assistant principal appears twice in the regression both as the effect of having an assistant principal in general and as the effect of the assistant principal as part of disciplinary resources, the interpretation of total effect of having an assistant principal on school quality must be taken in conjunction with both of these covariates. In other words, the total evaluation of assistant principal on reading quality must incorporate the partial effect of having an assistant principal plus the partial effect of having disciplinary resources per behavior problem. The total effect of an assistant principal can be expressed as follows:

$$\frac{dQ}{dAP} = \lambda_1 \frac{\partial Q}{\partial AP} + \lambda_2 \frac{\partial Q}{\partial Disc} \tag{14}$$

From the estimates in Table 3, this expression reduces to the following:

$$\frac{dQ}{dAP} = 3.41 + 24.28 \frac{1}{\text{\# of behvior problems per school}} \tag{15}$$

[26] Note that having an assistant principal was not a district requirement.

Figure 1 examines the total effect pertaining to having an assistant principal, with the measure of school quality in reading on the y-axis and number of behavior problems in a school along the x-axis. The graph highlights two points. First, a school with fewer behavior problems is associated with a higher school effect of having an assistant principal. In other words, as the number of behavior problems increases, the effect of having an assistant principal declines.

A second point stems from the fact that this analysis is based upon the theory of the education production function. In the production function literature, the law of diminishing returns states that as equal quantities of one variable (i.e., behavior problems) are increased while other factors remain constant (i.e., having an assistant principal), a point is reached beyond which the addition of one more unit of behavior problems results in a diminishing rate of return on the output (i.e., school quality). This is apparent in Figure 1. There is an initial steep decline in school quality with the first few increases in behavior problems. That is, there is a drop in school quality as the assistant principal must initially divest his disciplinary resources among behavior problems in the school. However, at such low levels of school counts of behavior problems, there is still a relatively high level of school quality associated with having an assistant principal. This pattern levels-out at high of behavior problems. That is to say, the differential effect of having an assistant principal blurs at high levels of behavior problems.

A final significant resource in the reading model is the effect of school nurses. Specifically, schools with nurses have higher reading quality than do schools without nurses ($\beta = 4.31$, $p<0.05$), holding all else equal. The standardized coefficient yields a result of 0.14σ. The final set of independent variables pertains to school environment. However, none are significant predictors of school quality in reading.

School Quality in Math

Table 4 presents parameter estimates, robust standard errors, and approximate *p* values from fitting the model in Equation (5) for math school quality. Note that measures of math quality were derived from SAT9 math achievement in Gottfried and Inman (2010). Similar to Table 3, there are two sections of results: one for unstandardized regression coefficients and one with standardized betas in order to interpret effect sizes.

To begin, schools with music programs also have higher math quality compared to schools that do not ($\beta = 4.78$, $p<0.05$). The standardized regression coefficient on this parameter suggests an effect size of 0.11σ. Unlike reading, having a language program is not related to school quality in math. However, similar to the results on reading, having an ELL program is not significantly related to school quality in mathematics.

Results from the set of variables pertaining to personnel resources indicate that holding all else equal, schools with nurses have higher school quality in math ($\beta = 7.20$, $p<0.01$) than do schools without school nurses. In terms of standardized betas, the associated effect size is 0.16 standard deviations. The result of having a school nurse is consistent with the results of the reading regressions, although the coefficient and standard error suggest a more highly significant effect in math.

Finally, results pertaining to school environment indicate that K-8 schools are negatively associated with elementary schooling quality in mathematics (β = -4.00, $p<0.05$). The associated standardized beta coefficient is -0.10σ. Other school-level environmental inputs are not significant, as consistent with reading.

Further Examining the Effect of "People"

Of the effect sizes portrayed in Tables 3 and 4, many of those that are larger in magnitude and that are statistically significant are found within the category of "people." This is evident across both reading and math regression models. Hence, there appears to be a consistent message in the analysis: out of all three categories of school-level variables, it may be the people at the institutional level seem to have a relatively larger relationship with school quality. Discerning how these people may influence their institutional environments is the focus of this final analytical section. In particular, the large effect size of having a school nurse in both reading and math analyses merits further investigation of the mechanism by which nurses can improve school quality.

This section examines two categories of people at the school level: disciplinarians and nurses. The analysis begins with an evaluation of the relationship between disciplinary resources and school quality. The results from Table 3 suggest a positive relationship between disciplinary resources and reading quality. It may be hypothesized that an increase in the breadth of disciplinary resources may diminish the number of behavior problems in a school. Fewer behavior problems, on average, may be related to higher classroom testing performance (Gottfried & Inman, 2010). This higher testing performance would then directly impact the quantifiable measure of school quality, as constructed in this study.

Table 5 presents the results of regressing the school average behavior grade on the independent variables from Table 1. Note that the behavior grade is on a scale of 1 (being a D – no F's are in the dataset) through 4 (being an A). Hence, a higher average behavior grade in a school implies fewer behavior problems in that given school.

Focusing on the estimates of disciplinarians in Table 5, the results suggest the following two interpretations. First, schools with assistant principals and police officers have lower average behavior grades, holding all else constant. The analysis here is not causal, and thus what this suggests, then, is logical: schools that have disciplinarians have them for good reason. A second interpretation, however, suggests insight into the relationship between disciplinarians and school quality. Looking at the number of disciplinarians per behavior problem (i.e., a measure of the breadth of disciplinary school resources), schools with more disciplinary resources per behavior problem tend to also have higher behavior grade averages, which recall is a positive school attribute in the coding of the behavior grade variable.

Thus, as disciplinary resources are made increasingly greater, their effectiveness also increases in raising the average level of behavior in their schools. In fact, the effect size suggests that a one standard deviation increase in disciplinary resources available at the school is related to a 0.27 standard deviation increase in average student behavior. Thus, if disciplinary resources relates positively to average student behavior, and if better student behavior relates to testing performance (Gottfried & Inman, 2009), then the relationship may suggest here that disciplinary resources may relate positively to school quality through its relationship with school behavior, since school quality is based upon student-level testing performance.

Research also suggests that schools with higher patterns of absences tend to also experience lower student performance on exams (Caldas 1993; Lamdin 1996). Thus, if lower levels of absences relate to higher testing performance, and if testing performance directly creates the measure of school quality as developed in this paper, then this might suggest that lower absences may relate to measures of school quality. Thus, understanding what factors may influence school absences can provide a more refined picture of the mechanism by which institutional-level variables relate to school quality.

In addition to presenting school behavior as an outcome, Table 5 also presents the results of regressing average school absence rate on the independent variables from Table 1. The results here are pertinent to both disciplinarians and nurses. To begin with disciplinarians, a similar interpretation is available here as for that in Table 5. Assistant principals and police officers are associated with higher school absence rates, holding all else constant. As before, the analysis here is non-causal, and thus the result is consistent with prior hypotheses: schools that require an assistant principal or police officer also have greater absences, as proxied by the dependent variable in this particular model.

The results pertaining to disciplinary resources are also consistent with depiction of how institutional-level factors relate to behavior as an outcome. Here, as schools with greater disciplinary resources have lower average rates of absences. The effect size is -0.19σ. Nurses also significantly relate to school absences: schools that have nurses tend to have lower average absence rates in a given school year. The size of the effect in Table 5 is -.19σ.

Thus, if disciplinary and nursing resources relate positively to lower school absences, and if the research suggests that fewer absences relate positively to higher testing performance (Dryfoos, 1990; Finn, 1993; Gottfried, 2009; Lehr et al., 2004; Stouthamer-Loeber & Loeber, 1988), then the relationship here indicates that having a larger breadth of disciplinary resources and having school nurses may relate positively to this study's measure of school quality, as quality is constructed based on student-level testing performance.

Discussion and Conclusion

This study has contributed to the research on school effectiveness. By first implementing the theory of educational production from the economics of education literature, this paper then empirically evaluated the model to determine which school-level factors relate to school quality, holding constant student, teacher, classroom, and neighborhood information. By conducting regression analyses on an institutional-level dataset of elementary schools within the District of Philadelphia over the years 1997 through 2000, this study has provided evidence that a range of school-level resources – as broken out by programs, people, and property – have significant relationships to school effectiveness in standardized testing.

The results indicate that there are significant relationships between school resources and quality in both reading and math testing subject areas. While the specific effects of the empirical analyses portray differential results depending on each testing outcome,[27] the results nonetheless indicate that all three categories of variables relating school quality are represented significantly across the analyses. Even though there may be distinctive results between reading and math, they nonetheless contribute to the overall analysis of school quality.

[27] A secondary analysis tested for interactions between variables in all three categories and the school's relative quality in the district, as determined by percentiles. However, no significant relationship existed for any interaction, not in reading or math.

The results of school-wide programming suggest a positive impact of language skills and music programs. Specially, schools with language skills programs have higher reading test effectiveness than do schools without such programs. Analogously, schools with music programs have higher effectiveness in math standardized test-taking than do school without music programs. These results are consistent with previous literature which has provided evidence of significant relationships between language skills programs and reading test performance (Ball & Blachman, 1988) and music programs and math test performance (Gardiner et al., 1996).

The analysis of school-level human resources indicates a positive relationship between the breadth of disciplinary resources and reading effectiveness, though no significant relationship exists for math. Gottfried and Inman (2010) provided causal evidence of a negative relationship between behavior problems and the classroom experience and subsequent test performance. The results here suggest that in addition to negative individual and classroom effects, there is also a third effect in play: an overall reduction in school quality from having a higher level of behavior problems, which as Table 5 suggested may be mediated through the breadth (or lack thereof) of disciplinary resources.

Consistently, for reading but not for math effectiveness, schools with assistant principals tend to have higher school quality, holding all else constant. Additionally, because the effect of having an assistant principal has been constructed as part of the indicator of a school's span of disciplinary resources, the results for reading indicate diminishing marginal returns on the total effect of a school's assistant principal. While still positive, at large levels of school behavior problems, the effect of having assistant principal as an input to the schooling quality education production function loses its potency. Nonetheless, the results correspond with much of the literature on school leadership and school effectiveness which finds significant relationships between having assistant principals as part of a larger administrative staff and subsequent school outcomes (Hallinger & Heck, 1998).

Additional analyses of school-level personnel indicate that, for both reading and math effectiveness, schools with nurses tend to have higher school quality. Much of the literature pertaining to health and education would find these results consistent. The research has suggested that upwards of 30 percent of children experience injuries around schools (Peterson, 2002). Thus, the impact of a having a school nurse has been shown to prevent health issues and injuries: when students can be treated on site, research suggests a subsequent decrease in health-related absences and an increase in classroom time and instruction (Allen, 2003; Guttu, Engelke, & Swanson, 2004). The results of Table 5 support this finding by demonstrating the extent to which nurses significantly relate to school absences.

Generally, school environmental resources do not indicate any particularly significant relationships to testing effectiveness in reading or math. There is one exception, however: schools which span kindergarten through 8^{th} grade have lower math school effectiveness than do schools which are strictly elementary. This result may seem contradictory with much of the literature, which has found positive effects on the testing outcomes of middle school students in K-8 schools compared to those in separated middle schools (Byrnes & Ruby, 2007; Coladarci & Hancock, 2002; Offenberg, 2001). However, those studies had solely evaluated the educational and psychological effects of K-8 schools on middle school educational outcomes, whereas the results here would suggest a negative effect for those elementary school students coupled in the same buildings as middle schoolers.

Because this study focuses on a particularly high-poverty and high-minority group of urban school children, the contributions of this paper extend beyond the empirical evaluation of the relationship between school-level inputs and school quality. Rather, this research has unified previous research by bringing to the foreground an array of institutional-level factors that have both positive and negative influences on the urban school experiences for at-risk youth in early years of education. Since the consequences of educational failure are exacerbated for children in large urban cities such as Philadelphia (Beaton et al., 1996; Byrnes & Ruby, 2007; Schmidt et al., 1999), having school-level information in addition to student- and classroom-level data yields insight into the resources that can improve the educational attainment of at-risk students in large urban cities.

What this paper has shown, then, is that even after accounting for student, teacher, classroom, and neighborhood data, school-level resources continue to impact the educational experiences of urban youth. Thus, by identifying those school-level factors that relate school quality to programs, personnel, and school environment, this study has demonstrated that particular institutional characteristics of urban elementary schools can significantly influence school effectiveness, above-and-beyond student or classroom circumstances. As such, the results of this paper can be used to better identify those institutional challenges faced by urban schools, how these challenges are actualized, and moreover, the type and level of resources necessary to reform schooling for at-risk youth.

This paper, furthermore, highlights the value of having detailed school-level data in determining relationships between institutional structures and characteristics and the outcomes of their students. By distinguishing among three particular categories of institutional-level resources – programs, people, and property – for each elementary school in the School District of Philadelphia, this paper demonstrates how academics, researchers, policy makers, and practitioners can disaggregate indicators of school quality into more useful metrics to better understand the channels through which school quality is affected. Identifying these factors has provided insight as to whether school quality matters and also points to which factors play a more significant role than others. As a result of doing so, this paper has not only enabled for a more detailed implementation of the education production function, but has also allowed for a more in depth understanding of how specific educational and financial resources can affect the urban school experience.

Further research can build upon the work in this paper in several capacities. For instance, while the data have been comprehensive in its scope of student, classroom, teacher, neighborhood, and school variables, there remains the opportunity to improve on information relating to management and leadership. Specifically, the data do not contain information on the specifics of school principals and thus the model could not parse out indicators relating to principal qualifications or managerial style. These variables remain in the error term of the analysis, and this may account for small R^2 values in regression tables. Thus, an extension of this research would link the data utilized in this paper to additional administrative data regarding principal qualifications as well as survey data containing principal and teacher reflections on concurrent school leadership. While cross-sectional surveys have been conducted around school leadership styles in the School District of Philadelphia (e.g., Tighe, Wang, & Foley, 2002), the longitudinal nature of this research study poses an additional challenge of linking measures of school management over time, since the fundamental goal of this paper has been to quantify year-specific school effectiveness.

Furthermore, the data used in this study are restricted to the analysis of elementary school outcomes. However, a longitudinal dataset that contains elementary, middle, and high school observations could provide insight on the relationship between early effects of schooling resources on future academic success. Recent increases in school accountability certainly provide research prospects of empirically unifying these different levels of education and across many levels of data – from individual students to institutional leaders.

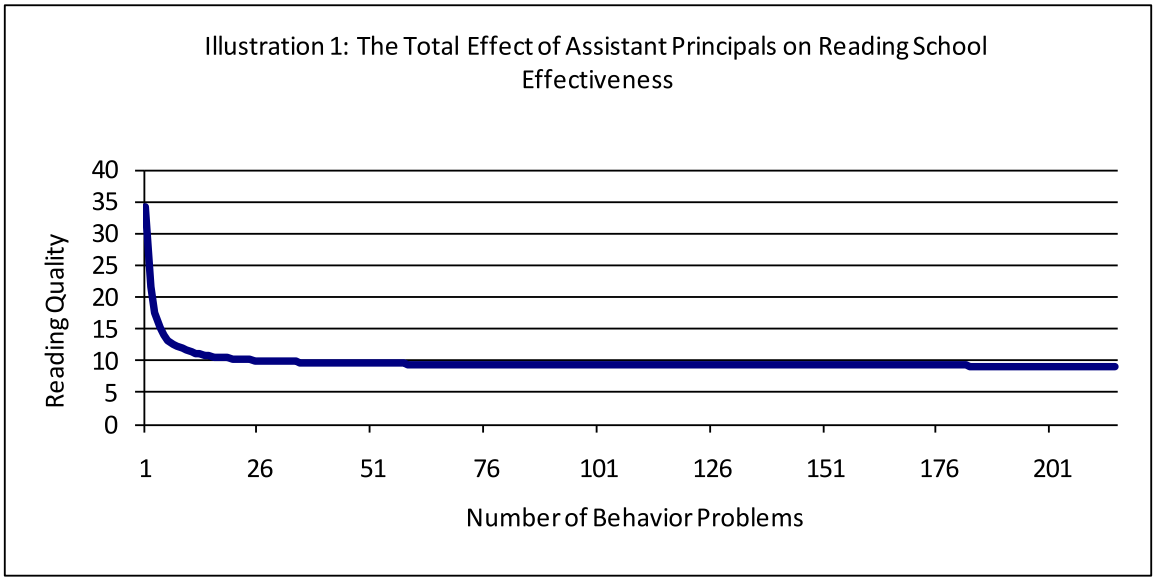
Illustration 1: The Total Effect of Assistant Principals on Reading School Effectiveness
Reading Quality
40
35
30
25
20
15
10
5
0
1
26
51
76
101
126
151
176
201
Number of Behavior Problems

Table 1

Descriptive Statistics of School-Level Production Function Inputs and Outputs

	Mean	*SD*
Outcomes		
Reading school quality	0.00	*11.83*
Math school quality	0.00	*17.53*
Inputs: Programs		
Music	0.25	*0.44*
Language	0.51	*0.50*
ELL	0.17	*0.38*
Inputs: People		
Special ed teachers per special ed student	0.11	*0.08*
Disciplinary resources per behavior problem	0.01	*0.05*
Assistant principal	0.12	*0.33*
Safety Officer	0.17	*0.37*
School community liaison	0.02	*0.13*
Nurse	0.25	*0.38*
Inputs: School Environment		
Teachers per student	0.03	*0.00*
Square footage per student (in ft)	12.29	*13.24*
K-8	0.25	*0.43*
N	674	

Table 2

Parameter Estimates of Logistic Regressions

	Music	Language	ELL	Special Ed Teachers[(a)]	Assistant Principal	Police Officer	Community Liaison	Nurse
School Academic Demographics (percent of student body)								
ELL	9.17	-12.03	-0.81	0.42	0.91	-0.52	0.37	1.66
	(9.05)	*(10.86)*	*(13.23)*	*(1.84)*	*(2.97)*	*(0.71)*	*(4.94)*	*(1.73)*
Special Education	-2.56	5.36	38.85	3.70	-2.36	-0.96	7.16	0.77
	(9.41)	*(15.01)*	*(39.74)*	*(2.36)*	*(4.81)*	*(0.91)*	*(7.62)*	*(0.95)*
Behavior problem	-4.45	-4.97	10.97	-2.70	3.15	-0.69	-3.04	-0.28
	(5.28)	*(6.38)*	*(8.05)*	*(1.09)*	*(2.10)*	*(0.42)*	*(4.60)*	*(0.44)*
Nonpromotion	30.48	5.02	70.62	8.34	-0.04	2.92	-105.22	2.71
	(36.15)	*(43.00)*	*(50.11)*	*(7.46)*	*(20.59)*	*(2.88)*	*(152.43)*	*(2.98)*
School Socioeconomic Indicators								
Free lunch recipients (percent)	0.69	3.82	-10.33	-0.36	-0.39	0.15	0.33	-0.11
	(3.24)	*(2.87)*	*(14.85)*	*(0.51)*	*(0.74)*	*(0.20)*	*(1.46)*	*(0.20)*
Average percent white of student census block	-9.21	2.74	-14.72	-1.26	-1.01	0.53	-1.01	-0.65
	(6.42)	*(9.11)*	*(13.72)*	*(1.49)*	*(0.74)*	*(0.57)*	*(1.31)*	*(0.59)*
Average income of student census block	2.83	-6.27	-1.61	-0.06	2.22	-0.06	3.40	0.37
	(4.43)	*(5.96)*	*(8.56)*	*(1.34)*	*(1.68)*	*(0.51)*	*(2.75)*	*(0.53)*
N	674	674	674	674	674	674	674	674
R^2	0.08	0.08	0.09	0.18	0.02	0.03	0.07	0.07

Notes: $^{***}p<0.01$, $^{**}p<0.05$, $^{*}p<0.10$.

(a) This regression is based on ordinary least squares, not logistic estimation.

Table 3

Regression Coefficients Predicting Reading School Effectiveness

	Unstandardized Coefficients		Effect Size	
	β	*SE*	β	*SE*
Inputs: Programs				
Music	-0.43	1.98	-0.02	0.07
Language	2.09 *	1.26	0.09 *	0.05
ELL	-0.22	1.71	-0.01	0.05
Inputs: People				
Special ed teachers per special ed student	4.95	6.93	0.03	0.05
Disciplinary resources per behavior problem	24.28 **	11.28	0.06 **	0.03
Assistant principal	3.41 *	1.98	0.11 *	0.06
Safety Officer	-0.20	1.93	-0.01	0.06
School community liaison	6.09	4.22	0.08	0.06
Nurse	4.32 *	2.35	0.14 *	0.08
Inputs: Environment				
Teachers per student	-108.70	263.68	-0.03	0.06
Square footage per student (in ft)	-0.05	0.04	-0.06	0.05
K-8	0.37	1.51	0.01	0.06
R^2	0.061		0.061	
N	392		392	

Notes: *** *p* <0.01, ***p* <0.05, **p* <0.10.

Table 4

Regression Coefficients Predicting Mathematics School Effectivenss

	Unstandardized Coefficients		Effect Size	
	β	*SE*	β	*SE*
Inputs: Programs				
Music	4.78 **	2.19	0.117 **	0.054
Language	1.47	1.76	0.042	0.050
ELL	-2.23	2.45	-0.047	0.052
Inputs: People				
Special ed teachers per special ed student	2.19	10.14	0.010	0.048
Disciplinary resources per behavior problem	-0.90	9.08	-0.002	0.016
Assistant principal	3.98	2.60	0.083	0.054
Safety Officer	1.35	2.24	0.030	0.050
School community liaison	3.08	6.43	0.027	0.057
Nurse	7.20 ***	2.69	0.163 ***	0.061
Inputs: Environment				
Teachers per student	147.98	326.25	0.023	0.052
Square footage per student (in ft)	0.01	0.06	0.007	0.055
K-8	-4.00 *	2.14	-0.102 *	0.055
R^2	478		478	
N	0.07		0.07	

Notes: *** *p* <0.01, ** *p* <0.05, * *p* <0.10.

Table 5

Regression Coefficients and Effect Sizes Predicting Average School Behavior and Average School Absence Rate

	Average School Behavior		**Average School Absence Rate**	
	Coefficient	Effect Size	Coefficient	Effect Size
Inputs: Programs				
Music	0.15 **	0.26 **	-0.01	-0.21
Language	0.17 ***	0.36 ***	0.00	0.03
ELL	0.05	0.08	0.00	-0.02
Inputs: People				
Special ed teachers per special ed student	-0.29	-0.09	0.03 *	0.16 *
Disciplinary resources per behavior problem	1.43 ***	0.27 ***	-0.06 **	-0.19 **
Assistant principal	-0.23 ***	-0.32 ***	0.02 ***	0.35 ***
Safety Officer	-0.23 ***	-0.36 ***	0.02 ***	0.40 ***
School community liaison	0.04	0.02	0.00	0.01
Nurse	-0.01	-0.02	-0.01 ***	-0.19 ***
Inputs: Environment				
Teachers per student	5.97	0.07	-0.38	-0.07
Square footage per student (in ft)	0.00	-0.01	0.00	0.13
K-8	0.20 ***	0.35 ***	0.00	-0.12
R^2	0.14		0.07	
N	578		580	

Notes: *** *p* <0.01, ***p* <0.05, **p* <0.10.

CHAPTER 3

ASSESSING ACCESS: IS THERE EQUITY WITHIN URBAN SCHOOLING?

Chapter Abstract[28]

Partitioning variance has been used extensively in educational research as a tool to determine possible sources of school-to-school variation. This paper contributes new insight by assessing if and why there is significant variation in standardized testing performance for entire populations of cohorts within all elementary schools in a single urban school district. Specifically, this study evaluates variance in SAT9 reading and math scores over four academic years and within three analytical levels of the educational experience – student, classroom, and school. To do so, this study employs three-level hierarchical linear modeling (HLM) to determine how the overall variance in testing performance can be partitioned within classrooms, between classrooms, and between schools. The initial results indicate that the overwhelmingly most significant contributor to total variance in achievement is within classrooms at the student level. However, incorporating variables into a three-tiered model of student achievement explains the majority of the between classroom and between school variance, though only half of the within classroom variance.

[28] This chapter represents collaborative research with Erica L. Johnson.

Introduction

Urban school districts face an ever-increasing number of problems. Students have high rates of truancy, low rates of graduation, and their achievement levels are less than those of their peers in non-urban settings (Tobin, Seiler, & Walls, 1999: Waxman & Padron, 1995). In addition, teachers in urban schools are more likely to be underprepared and have limited access to material resources (Clewell et al., 1995; Wykoff et al., 2002). To deal with large classes and little equipment, many urban teachers use whole-class instructional techniques (e.g., lectures, class reading, and completing worksheets) in which students are passive learners. This type of instruction was characterized by Haberman (1991) as a 'pedagogy of poverty,' in which there are few opportunities for developing higher-order thinking skills.

From this perspective, urban schools face an enormous challenge. They have fewer resources and more constraints than their non-urban school counterparts, and yet compounding these issues, they are in desperate need of improving student achievement. Urban school leaders themselves are aware of this dilemma – that academic achievement and teacher recruitment continue to suffer in their resource-constrained schools (Lewis, Baker, & Jepson, 2000). As alarming as this may be to researchers and practitioners, persistent and widespread differences continue to exist in the access, retention, and achievement of urban students within and between districts. Given these differences in access to educational success, gaining equity in achievement is a particularly urgent goal in urban districts across the United States.

A focus on educational equity is pertinent for a number of reasons. First, equalizing achievement levels has been a longstanding issue in education and policy ever since the *Coleman Report* (Coleman et al., 1966) documented differences in student performance; The study highlighted the particularly low achievement levels of socially disadvantaged black students as compared to the rest of the students. Since then, many state courts have intervened in the state education systems due to these inequity concerns that Coleman (1966) had brought to surface. For example, the long-running *Abbott v. Burke* cases in New Jersey focused on inequities in 31 urban school districts as compared to the rest of the state. The results of these cases found that the education provided to the students in poor, urban districts was inadequate:

> "A thorough and efficient education requires such level of education as will enable all students to function as citizens and workers in the same society, and that necessarily means that in poor urban districts something more must be added." -*Abbott v. Burke*, 1990

These cases mostly pertained to a focus on equalizing school finance and other institutional resources as a way of improving achievement levels. However, inequities in student-level outcomes are often viewed as symptoms of unequal financial as well as non-financial resources. Thus both inequities in schooling inputs and outputs merit further investigation.

Second, addressing the issue of equity is especially crucial for large urban districts in the United States, as they encompass approximately 25 percent of all school-age students, 25 percent of all poverty students, 30 percent of all English language learners, and nearly 50 percent of all minority children (Pew Charitable Trust, 1998). Urban schools have significantly more poverty students, English language learners, and minority students than the average public school (Jacob, 2007). In addition, there are potentially positive externalities to better urban education, such as lower crime rates and increased labor market participation (Lochner & Moretti, 2004), and increased civic participation and knowledge (Dee, 2004). Therefore, understanding sources of inequity within an urban district may allow researchers and practitioners to identify and eliminate those institutional practices that promote inequitable practices. For example, tracking and course selection are potential school policies that may promote or deter equity and could therefore account for substantially significant differences in schooling success (Guiton & Oakes, 1995). Other policies that could affect equity may include differential budgets, teacher hiring practices, and a variation in school-level resources.

Finally, much attention has been paid to differences in the quality of education in suburban versus urban school districts. However, differences within districts are often overlooked, particularly within urban districts. Comparing suburban schools to urban schools solely at an aggregated level of analysis can be problematic: urban and suburban schools are often analyzed as homogeneous samples within their respective category and are yet viewed as extremely heterogeneous between the two categories. However, it has not been made explicit how much heterogeneity exists within each respective category. Thus, this paper breaks down the differences within an urban school sample to analyze the sources of variation at a more refined level of analysis.

Within the general context of producing equitable educational outcomes and within the specific context of urban schooling, the goal of this study is to identify the degree of variability in achievement for the population of schools in a single district and further to identify how incorporating covariates into a model of student achievement can explain this variance. In particular, this study uses multilevel modeling to partition the variance across a comprehensive set of student, neighborhood, classroom, teacher, and school attributes that covary with reading and mathematics standardized testing achievement in the Philadelphia School District over a four year academic period. An understanding of how schools differ in terms of student, classroom, and school resources may be useful for addressing inequities at a range of levels within the educational experience.

Background

In identifying sources of variability in academic achievement, early literature placed an emphasis on the ability and socioeconomic backgrounds of students rather than on classroom or school-level factors. The seminal *Coleman Report* (Coleman et al., 1966) suggested that little variance in achievement could be attributed to schools. Rather, this study assigned the variance in student achievement across schools as a function of family background, i.e., a student-level trait. Moreover, between classroom variance was also simply the result of the socioeconomic backgrounds of other students in the class (i.e., the peer group) more than anything else. Jencks et al. (1972) corroborated Coleman's results, suggesting that school budgets, policies, and characteristics were "secondary or completely irrelevant" (Jencks et al., 1972, p. 256). Instead, variation in achievement output was attributed to student-level inputs.

These two early pieces provided a foundation for future work. With more sophisticated methodological techniques and larger datasets, recent research has attempted to be more precise in distinguishing which inputs cause differences in the variation of achievement. In particular, recent empirical studies have consistently focused on three distinct hierarchical levels of the educational experience in their analyses – student, classroom, and school. Thus, building upon early research which supported student-level factors as the drivers of achievement variance, more recent work has examined differences in the variation of achievement based on between classrooms factors (Scheerens, 1993; Scheerens et al., 1989) or resources across schools (Lamb, 1997; Mortimore et al., 1988; Nuttall et al., 1989; Smith & Tomlinson, 1989).

Several research studies have primarily focused on the relationship between classroom-level factors and variance in student achievement. Beaton and O'Dwyer (2002) compared the standardized achievement of eighth grade student across a representative sample of countries using the Third International Math and Science Study (TIMSS) data. They reported that while student-level differences accounted for the majority of variation in student achievement within all countries, classroom-level and school-level differences also accounted for more variation in student achievement than what Coleman (1966) had suggested. Their results showed that in the United States, student-level factors accounted for 49 percent of the variation, classroom-level factors accounted for 35 percent of the variation, and school-level factors accounted for 15 percent of the variation in TIMSS achievement scores. Second, Lamb and Fullarton (2001) employed a subsample of U.S. and Australian schools from the TIMSS study. They found that classroom differences accounted for approximately one-third of the total variance in achievement in the U.S. and one-quarter in Australia. Finally, Hay/McBer (2000) examined 80 schools and 170 teachers in the United Kingdom to determine the longitudinal impact of teachers on the growth of achievement. The study reported that approximately 30 percent of variance in achievement was attributable to the classroom level.

Other research has examined how school-level factors impact variance in achievement. As an example, Bosker and Witziers (1996) employed multilevel modeling on a meta-analysis on school effectiveness research. Their work suggested that institutional-level factors account for eight to ten percent of variation in student achievement. Lee and Smith (1997) also found between school variance attributable to school-level factors. Using data from NELS: 88, the study presented a non-linear relationship between high school size and achievement and that these effects were magnified as SES decreased.

Overall, these studies suggest that student background as well as classrooms and schools can potentially impact the variation in achievement outcomes. A range of studies have examined differential effects in various nationally and internationally representative samples. However, none has examined an entire population of schools within a single, urban school district. This study has the advantage of using longitudinal, comprehensive, and non-selective data for multiple of cohorts of students in all elementary schools in a single district. Thus, this paper can draw conclusions based on entire populations at many levels of analysis rather than relying on samples. Specifically, this study employs data for the elementary schools in the Philadelphia School District from the 1995-96 to 1998-99 academic periods to investigate how student-level factors, classroom environment, and school-level resources account for variation in standardized testing performance in reading and math.[29] To do this, this study partitions the variance by using multilevel modeling procedures to estimate the amount of variance that can be attributed to each of the three levels.

The focus on elementary students in a single urban school district is particularly compelling for three substantive reasons. First, most studies have relied on outcomes of students within samples of middle or high schools. However, this research examines the educational experiences of students before they enter into later grades where the probability of school failure or behavior problems becomes intensified (Alexander, Entwisle, & Horsey, 1997; Barrington & Hendricks, 1989; Lehr et al., 2004).

[29] If we refer to a single year, that is the fall of that school year. For instance, we will use 1995-96 and 1995 interchangeably to refer to that school year.

Consequently, this study can identify early significant relationships in the schooling experiences of urban elementary youth. Thus, with results pertaining to how achievement varies across and within schools during early education, it may be possible to develop specific district policies for these highly at-risk students in primary schooling grades before future consequences become exacerbated.

Second, previous research has often focused on a single testing outcome (i.e., TIMSS). However, this paper presents results for both reading and math Stanford Achievement Test Ninth Edition (SAT9). This approach is germane within this study's urban school sample of elementary school students, because it is particularly those minority and high-poverty students who fall behind in achievement beginning as early as fourth grade (Balfanz & Byrnes, 2006). Third, this study evaluates elementary school outcomes over multiple years rather than as a single cross-section, as conducted in previous research. Consistency in the results over time would suggest robustness in the findings.

Data

The analysis of variance in this study is facilitated by a comprehensive dataset of student, neighborhood, teacher, classroom, and school observations.[30] Student, teacher, classrooms, and school data were obtained from the School District of Philadelphia via the District's Office of Student Records and through the District's Personnel Office. Neighborhood data were obtained from the 2000 Census flat files at the census block

[30] The inputs selected in the full model as covariates are based on those used in the education production function literature. For example, see Summers and Wolfe (1977).

level. Neighborhood data relating to age, sex, households, families, and housing units were merged from the Census Summary File 1; additional social, economic, and housing measures were merged from Summary File 3. The data sample in Summary File 3 includes one in six households that received the long-form Census survey, whereas Summary File 1 measures are based on the full universe of responding households. The dataset encompass all elementary schools in the School District of Philadelphia over the academic years spanning 1995-96 through 1998-99. Over this time period, the sample contains 26,581 student observations across 174 schools with elementary grades, either K-5 or K-8.

Student Data

Table 1 presents student and neighborhood data employed in this study for the sample over all years of the dataset. For each student in a given academic year, basic information concerning personal characteristics such as date of birth, gender and race is augmented by a rich selection of independent variables in two categories. First, academic performance variables include current and one-year lagged normal curve equivalent (NCE) scores in math and reading from the Stanford Achievement Test Ninth Edition (SAT9) for grades 2 through 4.[31] Because NCE scores are vertically scaled, the achievement scores from different grades are directly comparable. Second, students are identified according to: special education status; English language learning status; free lunch recipiency; having a behavior problem; and an indicator for whether the student had been enrolled in

[31] The NCE is the generally preferred measurement for methodological reasons; it has statistical properties that allow for evaluating achievement over time (Balfanz & Byrnes, 2006). NCEs range in value from 1 to 99 and have a mean of 50.

kindergarten within the Philadelphia School District. For each student observation, school, grade, and room assignments are available in each academic year.

In addition, information was collected on a student's home address, including street number and name and zip code. The merging of neighborhood data was achieved by geo-coding each address to its longitude and latitude and then assigning each student to a census block group. In the absence of directly observed family information, the vector of neighborhood variables often serves as proxies for unobserved family characteristics (Hanushek, Kain, Markman, & Rivkin, 2003).

Classroom Data

Table 1 also presents corresponding teacher data for each student observation in the database. For every teacher, basic time invariant characteristics, including race and gender, are augmented by variables in three major categories. First, appointment date variables are used to create a measure of experience include: district appointment date; teaching seniority date; and present position appointment date. Second, an educational history variable includes an indicator for whether a teacher has a Master's degree based on the graduate school code and name in the file. Third, certification variables allow for an indicator for whether the teacher is state certified, based on having completed Pennsylvania state certification level I or level II.

Table 1 also describes data for each student's classroom context. Because each student observation includes the school, grade, and classroom assignment of the student in each academic year, there is sufficient information to assemble classroom-level variables, including class size, average ability[32], and head count of peer characteristics (i.e., number of special education students in a room).

School Data

Table 1 also presents three categories of school-level variables. First, there are several measures related to school-wide programming. As presented in Table 1, these programs include music, language skills, and English language learning programs (ELL). Each program variable is a binary indicator as to whether or not a school has a specific program. A second set of independent variables includes non-instructional personnel resources that pertain to student discipline, parental support, and community outreach. They include indicators as to whether a school, in a given year, had an assistant principal, school nurse, or school community liaison. A final set of covariates pertains to the general school physical environment, including a measure of the school's total student enrollment and a binary variable indicating if a school is K-8 or K-5.

Methods

The purpose of this paper is to assess the variability of testing performance within classrooms, between classrooms, and between schools. The basis of the analysis of

[32] Classroom average ability is defined for student i as the mean of the 1-year lagged test scores of all other students (excluding i) in the room. There is a separate average ability variable for reading and math, applied for each respective outcome.

variance lies in the decomposition of variation, or sum of squares for the mean (SS). In this study's context, a basic model of the total sum of squares in test scores, Y, assumes that variation in test scores can decomposed into three separate components, each representing between school, between classroom, and within classroom units:

$$SS_Y = SS_{\substack{between\\schools}} + SS_{\substack{between\\classrooms}} + SS_{\substack{within\\classrooms}} \tag{1}$$

where

$$SS_Y = \sum_k \sum_j \sum_i \left(Y_{kji} - \overline{Y}\right) \tag{2}$$

in which $\overline{Y}$ is the mean of test scores over the entire sample (i.e., the grand mean). The summations are over all students *i* in classrooms *j* in schools *k*. Exploring equation (1) in more detail, $SS_{\substack{between\\schools}}$ is the portion of sum of squares in test scores that can be attributed to the variation between schools. In other words, the between schools sum of squares represents the deviation in the means of each school from the overall sample mean. It can be expressed as follows:

$$SS_{\substack{between\\schools}} = \sum_k N_k \left(\overline{Y}_{k\cdot\cdot} - \overline{Y}\right) \tag{3}$$

where $\overline{Y}_{k\cdot\cdot}$ is the mean of test scores at the school-level and N_k is the number of schools. Analogously, $SS_{\substack{between\\classrooms}}$ is the portion of the total sum of squares in test scores that can be attributed to variation at the classroom level, i.e., between classrooms in a school. It represents the deviance of each classroom mean from the mean test score in its particular school. As an equation, it can be expressed as follows:

$$SS_{\substack{between \\ classrooms}} = \sum_k \sum_j N_{kj} \left(\overline{Y}_{kj\cdot} - \overline{Y}_{k\cdot\cdot} \right) \tag{4}$$

where $\overline{Y}_{kj\cdot}$ is the mean of test scores between classrooms within each school in the sample and N_{kj} is the number of classrooms per school. Finally, the portion of the total sum of squares in test scores that can be attributed to students within a given classroom can be expressed as follows:

$$SS_{\substack{within \\ classrooms}} = \sum_k \sum_j \sum_i \left(Y_{kji} - \overline{Y}_{kj\cdot} \right) \tag{5}$$

A generally appropriate and accepted approach for empirically undertaking the analysis of variance based on the above decomposition is hierarchical linear modeling (HLM) (Bryk & Raudenbush, 1992). HLM takes into account the nested structure of the data: in this instance, students within classrooms within schools. Students within the same classroom (and analogously, classrooms within the same school) have more homogeneous learning environments than students randomly selected from the district because they share certain characteristics (e.g., teachers, peer groups, neighborhood characteristics). Therefore, the test scores of these students are not completely independent. However, HLM corrects for the non-independence of observations. This procedure allows not only for modeling of outcomes at multiple levels (in this case, 3), but also for allocating the variance at each level while simultaneously controlling for the variance across levels.

Specifically, this study uses HLM to partition the variance of two standardized testing outcomes, SAT9 reading and math achievement scores, across three distinct levels: within classrooms (student level), between classrooms within schools (classroom level), and between schools (school level). This allows for the explicit partitioning of variance among the three levels across four years and two testing outcomes and for the determination of which hierarchical level is associated with the greatest source of variation in a given academic year. Moreover, the nature of panel data enables an evaluation of whether sources of variation change over time.

In addition, using HLM enables the variance to be partitioned under two different fixed coefficients modeling regimes for both reading and math outcomes – the "null" model, which includes no covariates (except for the identification of students, classrooms, and schools), and the "full" model which includes all inputs presented in Table 1. Implementing the null model is equivalent to one-way ANOVA and involves fitting a variance-components model to estimate the amount of unidentified variance due to the effects of students (level 1), classrooms (level 2), and schools (level 3). By then including a full range of covariates pertaining to each respective level, this study explores if the available variation at each of three levels can be explained by controlling for a rich set of student-, classroom-, and school-level factors that predict reading and math achievement.[33] In essence, the full model assesses how much of the variance is

---

[33] In the full model, continuous variables at the student and classroom level were centered around classroom and school means, respectively. Continuous variables at the school level were centered around the grand mean. Binary variables at each level remained uncentered. Group mean centering allows for the

associated with measures of student demographics, classroom learning environments, and institutional resources.

Results

Null Model

Figures 1 and 2 present the relationships between the means and standard deviations of reading and math achievement, respectively, for all schools in the sample over the entire time period of the dataset. In general, for values of school test performance on the x-axis, there is a spread of standard deviation values on the y-axis, indicating a large variation in the variance of student performance within the population of elementary schools in the district. The fact that there are large standard deviations for higher scoring schools is worth noting: even at these high performing schools, there oftentimes remains a large variance in individual student performance. In fact, only at those schools at the extremes – with very low or high mean performance levels – is there a smaller variance in achievement.

The results for the HLM analyses of variance in the null model for reading and math achievement outcomes are presented in Tables 2 and 3, respectively. These tables provide the partitioning of total variance in reading and math achievement into: between school, between classrooms within schools, and within classroom components. The table entries are percentages of the total variance at each hierarchical level. For a given year,

---

estimation of within-classroom, between classroom, and between school effects at each distinct level (Bryk & Raudenbush, 1992; Kreft & de Leeuw, 1998) the percentages in that row add up to 100 percent, and the first column in each table provides the total available variance in the sample of schools as defined by the HLM algorithm.

Beginning with the variance that is partitioned over the entire panel of data, the bottom row of Table 2 suggests that between school and between classroom variation each explain approximately 15 percent of the total available variance in reading outcomes.[34] In contrast, within classroom variation accounts for 70 percent of the total variance in reading scores. Hence, variation at the student level (i.e., within each classroom) accounts for the majority of total variance in the model. A similar interpretation is found for math in Table 3. In the final row of the table, which provides the partitioned variance over the time span of the entire dataset, between school and between classroom variance account for 16 percent and 11 percent, respectively, of the total variance. Within classroom variation explains 70 percent of the total variability in math achievement in the data, and this result is consistent with reading in Table 2.

For any individual year in both reading and math, the results suggest a similar interpretation to the overall trends. There is considerable within classroom variation in achievement, with much smaller variation partitioned between classrooms and between schools. Thus, these results indicate that the biggest contributor to the variance in test scores lies within classrooms in any given academic year. Specifically, in reading,

[34] A smaller variation at the classroom level than what has been seen in recent literature is attributed to the fact that this study's sample is a population of classrooms in the same district. Other analyses have relied on representative samples, thus creating a larger variation from such a diversity in rooms caused by increased heterogeneity.

between 1995 and 1998, within classroom variation explains between 68 percent and 76 percent of the total variance in reading achievement scores in the district. On the other hand, between classroom variation explains approximately 8 percent to 13 percent and between school variation explains 13 percent to 19 percent.

The variance in math achievement across the three levels of the null model demonstrates a similar interpretation to reading. Within classroom student level variation explains between 52 percent and 62 percent, between classroom variation explains between 16 percent and 25 percent, and between school variation explains between 20 percent and 23 percent. For math, less variation is explained at the within classroom level of the null model than for reading, whereas there is a slightly larger portion of variance explained at both classroom and school levels, i.e., between classrooms and between schools, respectively. Nevertheless, the results are generally consistent across both testing subject areas in any given year and across all years spanned in the dataset.

Full Model

To explain the sources of variance, a full model incorporates a wide range of covariates into the analysis. Including the variables from Table 1 into an expanded model explains between 49 percent and 69 percent of the total variance in reading scores and 54 percent to 62 percent of the total variance in math achievement, depending on the academic year. These results are presented in Appendix Table 1 for reading achievement and Appendix Table 2 for math. These tables are constructed analogously to Tables 2 and 3, respectively. As before, the biggest contributor to the variation in test scores remains within classrooms at the student level, and the overall partitioning of variance is consistent to the results of the null model.

The total variance associated with the parameters at student, classroom, and school levels (i.e., the full models) are presented for reading in Table 4. Examining the first column shows that over the span of the dataset, the total variance explained by the full model is the highest in 1998 (69 percent), though years 1995-1997 do not fall much farther behind. Note that the percentages of variance for each hierarchical level do not sum to 100 percent. Rather, the value in each column is the percentage of variance explained at each specific level in the hierarchical framework. For example, in 1995's reading achievement, approximately 65 percent of the originally partitioned variance between schools can now be explained by the full model. These results can be directly related to the null models in Tables 2 and 3. Using 1995 reading achievement again, 65 percent of the available 13 percent variance between schools, plus 31 percent of the available 13 percent variance between classrooms, plus 49 percent of the available 74 percent variance within classrooms compose the total percent of variance explained by the full model in 1995: 49 percent.

Over the time span of the entire dataset, approximately one-half of within classroom variance is explained by the set of covariates incorporated into the full model, with a slightly higher percentage accounted for in the final year of Table 4. In the null model, within classroom variance was attributed as providing the largest source of variance. However, the percentages in Table 4 indicate that the full model only explains about 50 percent of the within classroom variation, even though this full specification has incorporated measures of prior achievement as well as demographic, academic, and neighborhood information for each student. Thus, the largest source of variation in achievement remains only halfway explained, even with the incorporation of a broad set of student-level covariates.

On the other hand, the variables in the full model at the second hierarchical level, which pertain to classroom composition and teacher characteristics, have accounted for a fairly large portion of the available between classroom variance as described by the null model. For the majority of the panel in Table 4, the implementation of the full model can explain over 65 percent and up to 89 percent of the available between classroom variance. The covariates in the full model are also highly associated with the between school variance across the years of the dataset. Recall that the full model includes measures pertaining to school-wide programs, non-instructional human resources, and the school-level environment. Having incorporated these three categories of covariates suggests that up to 91 percent of the between school variance can be explained in the full model, depending on the academic year. Such high values for between classroom and between school percentages of variance explained implies that finding additional variables at these two levels that increase the predictive power of reading performance will prove to be challenging. Adding additional variables can only explain what residual variance is left, though by looking at the results for 1998, this is not much.

Table 5 presents results for math achievement. This table portrays a similar explanation for math as Table 4 did for reading. At the student level, approximately one-half of within classroom variance is explained by including the set of covariates in the model, with a slightly higher percentage in the final year of Table 5. As with reading achievement, this result here indicates that the wide range of variables within the full model can only explain around 50 percent of the total variation of math achievement at the student level, even though this specification accounts for prior SAT9 math performance as well as other student covariates.

Table 5 also suggests consistent between classroom and between school interpretations of the percent of variance explained for math achievement as for reading. As for variance between classrooms, the full model for mathematics achievement – which includes classroom composition and teacher characteristics – accounts for a fairly large portion of explained variance, ranging from 60 percent to 98 percent. Such high percentages for both reading and math indicate that within schools in Philadelphia, between classroom variance, though not responsible for a large proportion of total variance, can nonetheless be explained fairly substantially with the full model designed in this study.

The covariates in the full model are also highly associated with the between school variance across the years of the dataset. Incorporating these covariates indicates that a majority of the between school variance can be explained with the full model specified in this paper. As with between classroom variance, between school variance does not account for a large portion of total variance in reading or math, as described in Tables 2 and 3. However, the full model indicates that the covariates pertaining institutional resources can account for the majority of the variance at this educational level of analysis.

Discussion

This study contributes new insight into issues of equity within urban schooling. By partitioning the variance components of the elementary educational experience into three levels – student, classroom, and school – this paper has presented empirical evidence on two questions: first, what is the source of the variance in achievement within a single school district and second, what explains it? In other words, this study has evaluated which levels of urban schooling contribute to differentiation in academic performance, and then which factors can help explain these differences. Because educational equity is an ongoing concern among practitioners and researchers, these results are an important step in understanding which student characteristics, classroom factors, and school resources can undermine equal access.

The focus on a single urban school district over time has enabled this study to document the patterns of variance as cohorts of students progress through early years of schooling. The analysis has demonstrated that not only is there variance in achievement within a single district, but that consistent patterns of variance are evident across multiple levels of the educational experience and across several years of data. Thus, in conjunction with previous studies, this paper also supports the premise that a significant relationship exists between student-, classroom-, and school-level factors and the subsequent variance in achievement.

Moreover, the evaluation of two standardized testing outcomes for elementary school students suggests that factors of variance have an impact across two testing subjects, both reading and math. The consistent results for both outcomes suggest there is a generalizability across multiple indicators of academic success for elementary school students. The results, across two measures and four academic years, thereby suggest a robustness in the findings.

Overall, in all years of the dataset, the null models (as well as the appendices) suggest that the overwhelmingly most significant contributor to total variance in achievement is at the student level: across both reading and math testing subjects, within classroom variance accounts for approximately two-thirds to three-fourths of the total variance in scores within the Philadelphia School District. Thus, within classrooms, student achievement is quite heterogeneous. On the other hand, the low percentages of classroom and school level variation indicate that the schools within the Philadelphia school district have similar average scores and that the classrooms within those schools are as well quite similar. These results suggest that the differences in resources at the school level contribute only a small amount to the variation in scores. Likewise, classroom variation makes a small contribution to variation in achievement. Instead, as first documented by the *Coleman Report* (Coleman et al, 1966), much of the variation in achievement remains at the student level.

Interestingly, the contribution of the covariates in the full model explain the majority of the between classroom and between school variance, though only half of the student-level variance. Thus, a significant portion of the variance at the student level remains to be explored, even after controlling for those factors of the full model, such as lagged achievement and other academic, demographic, and neighborhood characteristics. What's missing from the evaluation is directly observed attributes of student motivation and family characteristics and parental background.

It is particularly striking that over forty years since the publication of the *Coleman Report* (Coleman et al., 1966), this paper presents similar results. The overwhelmingly largest contributor to variation in achievement is at the student level and the rich set of covariates only explains about half of the variance at the student level. These two findings affirm those of the *Coleman Report*: family characteristics - outside the role of the school – may play a large, if not the most predominant, role in student achievement.

In conjunction with this latter point, there are three research extensions from this paper. First, because the data do not contain direct measures of family environments, it would also be beneficial to incorporate longitudinal metrics of parental involvement and engagement. In addition, having student psychological measures would yield greater insight into what drives variation in reading and math achievement. Second, the data used in this study were for elementary school students. A longitudinal dataset that contains elementary, middle, and high school observations could provide insight on the association between multilevel covariates and variance in achievement.

Third, it should be noted that this paper has focused on a single urban school district. While there are many advantages to evaluating a population of students within a single, large urban school district, it is possible that different results and interpretations may be found within other urban school districts or within districts that are suburban or rural. Also, it is possible that school-level resources could play a bigger role in the overall variation in test scores if those schools in the dataset had more heterogeneous resources, such as a statewide evaluation of urban, suburban, and rural schools in its domain. The results in this paper could then be compared to those using data from additional districts of different urbanicity in order to derive broader conclusions. Nonetheless, by employing multilevel methods and evaluating an entire population of urban schools in Philadelphia, this study has contributed new conclusions and built new foundations upon which we can evaluate issues of equity in education.

Illustration 1: Reading Achievement for Schools in the Philadelphia School District Sample, 1995-96 through 1998-99

Illustration 2: Math Achievement for Schools in the Philadelphia School District Sample, 1995-96 through 1998-99

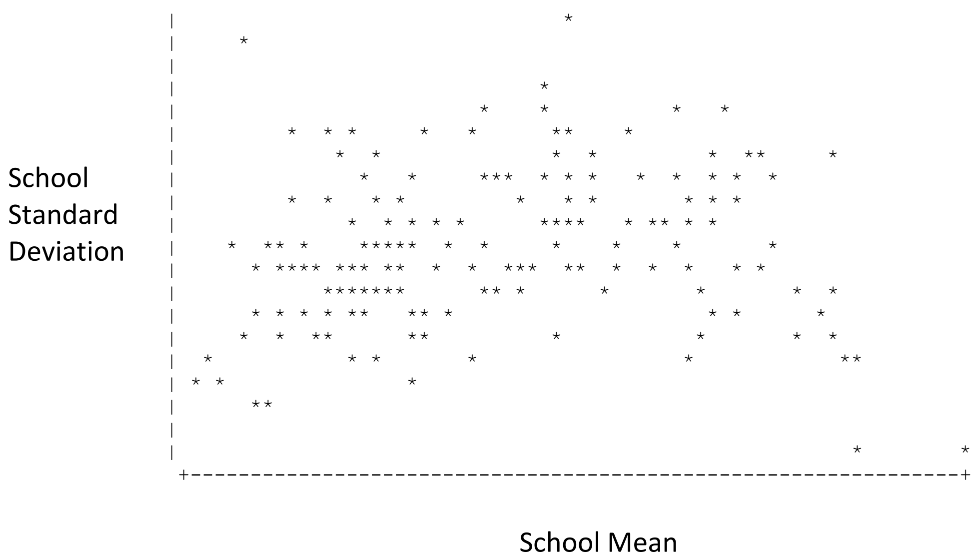

Table 1

Student, Classroom, and School variables

STUDENT LEVEL	*CLASSROOM LEVEL*
Demographic indicators	**Teacher variables**
Male	Male (binary)
Black	Years of experience
White	
Latino	**Classroom composition**
Asian	Class size
Other	Mean class reading score
Free lunch recipient	Mean class math score
	Class counts of:
	Behavior problems
Academic Indicators	ELL students
Special education	Special education students
ELL	Free lunch recipients
Attendend kg in district	Females
Lagged behv = A	
Lagged behv = B	*SCHOOL LEVEL*
Lagged behv = C	**School program indicators**
Lagged behv = D	Music program
	Language skills
Test Performance	ELL
SAT9 reading	
SAT9 math	**Personnel Resources Indicators**
Lagged SAT9 reading	Assistant principal
Lagged SAT9 math	School nurse
	School community liaison
Student neighborhood	
Block % white	**School Environment**
Block % poverty	Total enrollment
Block % vacant	Square footage
Avg block income (log)	K-8 (vs K-5) indicator

Table 2

Partitions of Variance, Reading Achievement: Null Model

		Reading: Percent of Variance		
Year	Total Available Variance	Between Schools	Between Classrooms	Within Classrooms
1995	231.56	13.25	12.52	74.24
1996	232.92	18.09	11.12	70.79
1997	212.78	19.03	13.37	67.59
1998	164.64	15.48	8.42	76.00
All Years	244.76	15.06	14.70	70.24

Table 3

Partitions of Variance, Math Achievement: Null Model

		Math Percent of Variance		
Year	Total Available Variance	Between Schools	Between Classrooms	Within Classrooms
1995	359.37066	15.14	15.58	69.27
1996	385.5545	20.80	13.98	65.22
1997	375.76405	21.45	14.29	64.26
1998	341.1804	19.38	16.68	63.94
All Years	377.62691	16.40	11.84	71.76

Table 4

Partitions of Variance, Reading Achievement: Full Model

		Reading: Percent of Variance Explained		
Year	Total Percent of Variance Modeled	Between Schools	Between Classrooms	Within Classrooms
1995	48.77	65.37	31.12	48.78
1996	52.26	54.12	79.28	47.54
1997	52.23	56.17	66.29	47.64
1998	69.24	91.21	88.82	62.83
All Years	54.51	56.60	64.76	51.91

Table 5

Partitions of Variance, Math Achievement: Full Model

		Math Percent of Variance Explained		
Year	Total Percent of Variance Modeled	Between Schools	Between Classrooms	Within Classrooms
1995	53.86	71.15	62.99	48.03
1996	55.56	56.10	98.12	46.26
1997	58.02	62.99	98.09	47.45
1998	61.96	70.78	59.16	60.02
All Years	55.12	58.37	80.83	50.14

APPENDICES

APPENDIX A

Variable Descriptions for Analyses in Chapter 1

Name	**Description**
Depenent variables	
SAT9 reading	Stanford Achievement Test, Ninth Edition
SAT9 math	Stanford Achievement Test, Ninth Edition
Student characteristics	
SAT9 reading lag	Previous year's SAT9 reading test score
SAT9 math lag	Previous year's SAT9 math test score
Lag behavior: A	Previous year's report-card behavior score is an A
Lag behavior: B	Previous year's report-card behavior score is a B
Lag behavior: C	Previous year's report-card behavior score is a C
Lag behavior: D	Previous year's report-card behavior score is a D
Male	Student is male
White	Student is white
Black	Student is black
Latino	Student is latino
Asian	Student is asian
Other	Student is other (i.e., pacific islander)
Repeat	Student has repeated current grade
Young	Student is young for his/her classroom. Young is determined by being less than the average age in the room and/or having a birthday in the following year's fall semester
K in Philadelphia	Student attended kindergarten within the Philadelphia School District
Special ed	Student is special education
Free lunch	Student is under the free or reduced lunch program
English language learner	Student is an english language learner
Classroom variables	
Class size	Count of students in a classroom
Class size^2	Squared value of Class Size
Mean reading	Mean of lagged SAT9 reading test scores in a classroom
Mean reading^2	Squared value of mean reading
Mean math	Mean of lagged SAT9 math test scores in a classroom
Mean math^2	Squared value of mean math
Mean x reading lag	Interaction of "Mean reading" with "SAT9 reading lag"
Mean x math lag	Interaction of "Mean math" with "SAT9 math lag"

APPENDIX A (cont'd)

Variable Descriptions for Analyses in Chapter 1

Name	**Description**
Peer variables	
Count of free lunch	Count of free or reduced lunch students in a classroom
Count of behv problems	Count of misbehaved students in a classroom
Count of ELL	Count of english language learner students in a classroom
Count of special ed	Count of special education students in a classroom
Count of female	Count of female students in a classroom
Teacher Characteristics	
Male	Teacher is male
White	Teacher is white
Black	Teacher is black
Hispanic	Teacher is hispanic
Asian	Teacher is asian
Other	Teacher is other (i.e., pacific islander)
Experience	Teacher's experience in the Philadelphia School District (in years)
Experience^2	Square of T. Experience
Certification	Teacher has Pennsylvania Instructional Level II certification
Masters	Teacher holds a Masters degree
Neighborhood characteristics	
Percent white	Percent of homes on student's census-level block that are white
Percent poverty	Percent of homes on student's census-level block that are below the poverty line
Log(income)	Log of median income on student's census-level block
Household vacancy	Percent of vacant homes on student's census-level block

APPENDIX A (cont'd)

Variable Descriptions for Analyses in Chapter 1

Late student regressions	
Class-size: non-late	Count of students in classroom, excluding those students arriving after the start of the academic year
Class-size: non-late^2	Squared value of Class-size (non-late)
Late arrivals	Count of students in classroom who have arrived at irregular times in the school year (i.e., not at the beginning of the year)
Count of free lunch: non-late	Percentage of free or reduced lunch students in a classroom, excluding the late arrivals
Count of behavior problem: non-late	Percentage of misbehaved students in a classroom, excluding the late arrivals
Count of ELL: non-late	Percentage of english language learner students in a classroom, excluding the late arrivals
Count of special ed: non-late	Percentage of special education students in a classroom, excluding the late arrivals
Count of female: non-late	Percentage of female students in a classroom, excluding the late arrivals
Count of free lunch: Late	Percentage of late-arriving free or reduced lunch students in a classroom
Count of behavior problems: late	Percentage of late-arriving misbehaved students in a classroom
Count of ELL: late	Percentage of late-arriving english language learner students in a classroom
Count of special ed: late	Percentage of late-arriving special education students in a classroom
Count of female: late	Percentage of late-arriving female students in a classroom

APPENDIX B

Grades-By-Schools Removed from the Trimmed Sample

Sch Id	Year	Gd	Sch Id	Year	Gd	Sch Id	Year	Gd
120	1995	2	147	1996	3	245	1997	3
120	1996	3	147	1996	4	245	1999	4
120	1998	3	147	1997	3	247	1996	3
120	1999	3	147	1997	4	247	1996	4
121	1996	3	147	1998	4	247	1997	3
121	1997	4	149	1997	4	247	1997	4
121	1998	4	149	1999	4	247	1998	4
123	1996	3	153	1996	4	248	1996	3
123	1998	4	153	1997	4	248	1996	4
125	1995	2	153	1998	4	249	1995	2
125	1997	3	153	1999	4	249	1996	3
125	1997	4	219	1997	4	249	1997	4
125	1999	3	220	1996	3	249	1998	4
126	1996	3	220	1996	4	251	1997	4
127	1997	3	220	2000	4	251	1999	4
127	1997	4	226	1995	2	252	1998	4
129	1998	4	226	1997	3	254	1996	4
130	1995	2	226	1998	4	254	1997	4
130	1996	4	226	1999	4	254	1998	4
131	1996	3	226	2000	4	254	1999	3
131	1997	4	230	1995	2	258	1996	4
131	1998	4	232	1995	2	258	1997	4
133	1996	3	232	1996	3	258	1999	4
133	1997	4	232	1996	4	263	1996	3
133	1999	3	232	1997	3	264	1995	2
134	1995	2	232	1997	4	264	1996	3
134	1996	3	232	1998	4	264	1996	4
134	1997	3	232	1999	3	267	1997	4
134	1997	4	232	2000	4	269	1995	2
134	1998	4	234	1996	3	269	1998	3
134	1999	4	234	1996	4	269	2000	4
134	2000	4	234	1997	3	272	1995	4
135	1996	3	234	1997	4	272	1996	3
135	1997	4	237	1996	3	272	1997	4
135	1998	4	237	1996	4	272	1998	4
136	1998	3	237	1998	3	272	1999	3
136	1998	4	237	1998	4	272	1999	4
137	1996	3	237	2000	4	272	2000	2
137	1997	3	239	1997	3	273	1996	3
137	1997	4	239	1997	4	421	1996	3
137	1998	4	239	1998	4	421	1996	4
138	1998	3	239	2000	4	421	1997	4
138	1998	4	242	1996	3	421	2000	4
139	1998	4	242	1997	3	424	1997	4
140	1999	4	242	1998	3	424	2000	4
140	2000	4	242	1999	4	426	1997	4
142	1997	4	242	2000	4	426	1998	3
143	1996	3	244	1996	3	426	1999	4
143	1999	4	244	1997	4	426	2000	4
146	1996	3	244	1998	4	428	1996	3
146	1997	4	244	1999	3	428	1997	4

APPENDIX B (cont'd)

Grades-By-Schools Removed from the Trimmed Sample

Sch Id	Year	Gd	Sch Id	Year	Gd	Sch Id	Year	Gd
429	1998	4	522	1997	4	541	1996	4
429	1999	4	522	1998	4	541	1997	3
430	1996	3	525	1997	4	541	1997	4
430	1996	4	525	1998	4	541	1998	3
430	1997	4	525	1999	3	541	1998	4
431	1995	2	526	1996	3	541	1999	3
431	1996	3	526	1996	4	541	1999	4
431	1997	4	526	1997	3	541	2000	4
431	1998	3	526	1997	4	542	1995	2
434	1996	3	526	1998	3	542	1996	3
434	1997	3	526	1998	4	542	1997	3
437	1996	3	526	1999	3	542	1997	4
437	2000	4	526	1999	4	542	1998	4
438	1997	4	526	2000	4	542	1999	4
438	1998	4	528	1997	4	542	2000	4
438	1999	4	528	1998	4	544	1995	2
439	1996	4	528	1999	4	544	1998	4
440	1995	2	529	1995	2	544	1999	3
440	2000	4	529	1996	3	544	1999	4
443	1995	2	529	1996	4	544	2000	4
443	1996	3	529	1997	3	547	1996	3
443	1996	4	529	1997	4	547	1996	4
443	1997	3	529	1998	4	547	1997	3
443	1997	4	529	1999	3	547	1997	4
443	1998	4	529	1999	4	547	1998	4
443	1999	4	530	2000	3	547	1999	3
445	1996	3	530	2000	4	547	1999	4
445	1997	3	531	1996	3	547	2000	4
447	1997	3	532	1997	3	548	1997	4
447	1999	3	532	1997	4	549	1995	2
447	2000	4	532	1998	3	549	1996	4
451	1995	2	532	1998	4	549	1998	4
451	1996	3	534	1996	4	549	1999	3
451	1997	4	534	1997	3	549	2000	4
451	1998	3	534	1997	4	550	1997	3
453	1997	3	537	1996	3	553	1998	3
453	1998	4	537	1997	4	553	1998	4
456	1996	3	537	1998	4	553	2000	4
456	1997	3	539	1995	2	556	1996	3
456	1997	4	539	1996	3	559	1996	4
457	1996	3	539	1996	4	559	1997	3
457	1997	3	539	1997	3	559	1997	4
457	1998	4	539	1997	4	559	1998	3
520	1996	3	539	1998	3	559	1998	4
520	1997	4	539	1998	4	559	1999	3
521	1997	4	539	1999	3	559	2000	4
521	1999	4	539	1999	4	568	1997	3
522	1995	2	539	2000	3	568	1998	3
522	1995	3	539	2000	4	568	1998	4
522	1996	3	540	1996	4	568	1999	4
522	1996	4	541	1996	3	620	1997	4

APPENDIX B (cont'd)

Grades-By-Schools Removed from the Trimmed Sample

Sch Id	Year	Gd	Sch Id	Year	Gd	Sch Id	Year	Gd
621	1997	4	644	1998	4	739	1997	4
621	2000	4	644	1999	4	740	1996	3
622	1996	4	645	1996	4	740	1996	4
622	1998	4	647	1996	3	740	1997	3
622	1999	3	647	1996	4	740	1999	3
622	2000	4	647	1997	4	740	1999	4
623	1996	3	647	1998	4	742	1998	3
623	1997	3	647	1999	4	742	2000	4
623	1999	3	720	1995	2	744	1996	4
624	1996	4	720	1996	3	744	1997	3
624	1998	4	720	1997	4	744	1997	4
624	2000	4	720	1999	3	744	1998	3
625	1997	3	721	1996	3	744	1998	4
625	1997	4	721	1997	3	744	1999	3
626	1996	3	721	1997	4	744	2000	4
626	1997	4	721	1999	3	746	2000	4
626	1998	4	722	1995	2	747	1996	3
626	2000	4	722	1996	3	747	1998	4
628	1997	4	722	1996	4	749	1996	3
628	1998	4	722	1997	3	749	1996	4
628	1999	4	722	1997	4	749	1997	3
628	2000	4	724	1996	3	749	1998	4
630	1997	4	724	1998	3	749	1999	4
630	1998	3	724	1999	3	749	2000	4
630	1998	4	724	1999	4	751	1996	3
630	1999	4	725	1999	3	751	1998	4
631	1998	4	726	1996	3	751	1999	3
631	2000	4	726	1998	4	753	1996	3
632	1997	4	726	1999	3	753	1998	3
633	1995	2	726	2000	4	753	1998	4
633	1996	3	728	2000	4	753	2000	4
633	1996	4	729	1996	3	818	1997	4
634	1995	2	729	1998	4	820	1998	4
634	1996	3	729	2000	4	820	1999	3
634	1997	4	730	1995	2	821	1996	3
634	1998	4	730	2000	4	821	1998	4
634	1999	3	731	1999	4	821	1999	3
634	1999	4	731	2000	4	821	2000	4
634	2000	4	732	1995	2	823	1996	3
635	1995	2	732	1998	4	823	1998	4
635	1996	3	733	1996	4	823	1999	3
635	1997	4	733	1997	3	824	1998	3
635	2000	4	733	1998	4	824	1998	4
638	1997	3	735	1998	3	824	1999	3
639	2000	4	735	1998	4	824	1999	4
643	1997	4	735	1999	4	825	1995	2
643	1998	4	735	2000	3	825	1996	3
643	1999	3	735	2000	4	825	1997	3
643	2000	4	738	1997	3	826	2000	4
644	1996	3	739	1996	4	827	1995	2
644	1997	4	739	1997	3	827	1996	3

APPENDIX B (cont'd)

Grades-By-Schools Removed from the Trimmed Sample

Sch Id	Year	Gd
827	1996	4
827	1997	3
827	1999	4
830	1997	4
831	1996	3
831	1997	4
831	1998	4
831	1999	3
831	2000	4
834	1997	4
835	1996	4
835	1997	3
835	1998	4
836	1996	3
836	1997	4
837	1996	3
837	1996	4
837	1998	4
839	1997	4
841	2000	4
842	1997	4
842	1999	3
843	1996	3
844	1999	3
844	2000	4

APPENDIX C

Effect Sizes for Observable Peer Characteristics

	Reading				Math			
	Baseline	Trimmed	Free Lunch	Non Free Lunch	Baseline	Trimmed	Free Lunch	Non Free Lunch
Count of free lunch	-0.070 ***	-0.065 *	-0.089 *	-0.056	-0.069 **	-0.071 **	-0.053	-0.086 **
Count of behavior problems	-0.033 ***	-0.026 ***	-0.031 *	-0.023	-0.024 **	-0.028	-0.047 **	-0.017
Count of ELL	-0.010	-0.014	-0.025	-0.006	-0.005	-0.014	-0.029	0.000
Count of special ed	-0.014	-0.019	-0.013	-0.026 *	-0.016 *	0.008	0.004	0.011
Count of females	0.055 ***	0.037 **	0.033 *	0.040 **	0.032 ***	0.033 **	0.030	0.034 **

*Note: *** p<.01; **p<.05; *p<.1*

Appendix D

Descriptive Statistics of Student Sample from Gottfried & Inman (2010)

	Mean*	SD
N	97,007	
SAT9 achievement outcomes		
Reading	39.11	15.36
Math	55.90	19.00
Reading, lagged	37.13	15.57
Math, lagged	56.92	18.59
Student race, in percent		
White	16.82	37.41
Black	67.85	46.70
Hispanic	10.99	31.28
Asian	4.17	19.98
Other	0.16	4.02
Student gender, in percent		
Male	48.90	49.99
Female	51.10	49.99
Academic indicators in percent		
Attended Phila kindergarten	85.14	35.57
Free lunch eligible	52.46	49.94
English language learner	3.65	18.76
Special education	4.04	19.68
Lagged behavior = D	10.54	30.71
Lagged behavior = C	22.83	41.97
Lagged behavior = B	35.29	47.79
Lagged behavior = A	31.34	46.39
Student's census block		
Block percentage: white	29.39	32.50
Block percentage: poverty	14.35	8.67
Block percentage: house vacancy	12.95	9.38
Log of income (in dollars)	10.15	0.45

*Note: Population is based on having observations with required test scores. Research sample is based on test scores and non-missing information for required independent variables.

Appendix D (cont'd)

Descriptive Statistics of Student Sample from Gottfried & Inman (2010)

Teacher race, in percent		
White	82.74	37.79
Black	16.26	36.90
Hispanic	0.68	8.21
Asian	0.28	5.26
Other	0.05	2.20
Teacher gender, in percent		
Male	7.88	26.94
Female	92.12	26.94
Teacher skills		
Teacher experience (in years)	3.81	7.70
Teacher state certified (percent)	24.23	24.23
Teacher has a masters degree (percent)	33.84	33.84
Class size (head count)	28.23	3.80
Academic classroom characteristics		
Mean SAT9 reading score	33.61	11.35
Mean SAT9 math score	53.69	12.46
Other classroom characteristics (head count)		
Free lunch	12.62	7.54
Behavior problems	1.62	1.75
English language learners	1.33	3.16
Special education	1.03	1.54
Female	10.37	3.62

*Note: Population is based on having observations with required test scores. Research sample is based on test scores and non-missing information for required independent variables.

Appendix E

Partitions of variance, reading achievement: Full model

		Percent of Variance		
Year	Total Percent of Variance Modeled	Between Schools	Between Classrooms	Within Classrooms
1995	48.77	17.76	7.99	74.25
1996	52.26	18.74	16.87	64.40
1997	52.23	17.88	20.47	61.66
1998	69.24	20.16	10.80	69.04
All Years	54.51	15.64	17.47	66.90

Appendix F

Partitions of variance, math achievement: Full model

		Percent of Variance		
Year	Total Percent of Variance Modeled	Between Schools	Between Classrooms	Within Classrooms
1995	53.86	20.00	18.22	61.77
1996	55.56	21.00	24.70	54.30
1997	58.02	23.29	24.15	52.56
1998	61.96	22.14	15.93	61.94
All Years	55.12	17.36	22.17	65.27

BIBLIOGRAPHY

Abbott v. Burke II. 119 N.J. 287, 575 A.2d 359 (1990).

Alexander, K.L., Entwisle, D.R., & Horsey, C.S. (1997). From first grade forward: Early foundation in high school dropout. *Sociology of Education*, *70,* 87-107.

Allen, G. (2003). The impact of elementary school nurses on student attendance. *The Journal of School Nursing, 19*, 225-231

Ammermuller, A., & Pischke, J.S. (2006). Peer effects in European primary schools; Evidence from PIRLS. Mahneim: ZEW Discussion Paper No. 06-027.

Angrist, J.D., & Lang, K. (2004). Does school integration generate peer effects? Evidence from Boston's Metco program. *American Economic Review, 94,* 1613-1634.

Argys, L.M., Rees, D.I., & Brewer, D.J. (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management*, *15*, 623- 645.

Arnott, R., & Rowse, J. (1987). Peer group effects and educational attainment. *Journal of Public Economics*, *32*, 287-305.

Balfanz, R., & Byrnes, V. (2006). Closing the mathematics achievement gap in high poverty middle schools: Enablers and constraints. *Journal of Education for Students Placed at Risk*, *11*, 143-159.

Ball, E., & Blachman, B.A. (1988). Phoneme segmentation training: effect on reading readiness. *Annals of Dyslexia*, *38*, 208-22.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice Hall: Englewood Cliffs, N.J.

Barrington, B., & Hendricks, B. (1989). Differentiating characteristics of high school graduates, dropouts, and nongraduates. *Journal of Educational Research*, *82*, 309-319.

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, M.A.: TIMSS Study Center.

Beaton, A.E., & O’Dwyer, L.M. (2002). Separating school, classroom, and student variances and their relationship to socio-economic status. In D.F. Robitaille & A.E. Beaton (Eds.), *Secondary Analysis of the TIMSS Data*. Netherlands: Kluwer Academic Publishers.

Betts, J.R. (1995). Does school quality matter? Evidence from the National Longitudinal Survey of Youth. *Review of Economics and Statistics*, *77*, 231-250.

Betts, J.R., & Zau, A. (2004). Peer groups and academic achievement: Panel evidence from administrative data. unpublished manuscript

Borman, G.D., Slavin, R.E., Cheung, A.C.K., Chamberlain, A.M., Madden, N.A., & Chambers, B. (2005). The national randomized field trial of Success for All: Second-year outcomes. *American Education Research Journal*, *42*, 673-696.

Bosker, R., & Witziers, B. (1996). *The magnitude of school effects or does it really matter which school a student attends?* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Byrnes, V., & Ruby, A. (2007). Comparing achievement between K-8 and middle schools: A large scale empirical study. *American Journal of Education*, *114*, 101-135

Caldas, S. J., & Bankston III, C. (1997). Effect of school population socioeconomic status on individual academic achievement. *Journal of Educational Research*, *90*, 269-277.

Chambers, J., & Parrish, T. (1992). *Meeting the challenge of diversity: An evaluation of programs for pupils with limited proficiency in English. Vol. IV, cost of programs and services for LEP students*. Berkeley, CA: BW Associates.

Clewell, B.C. (1995). *Systemic reform in mathematics and science education: An urban perspective*. Washington D.C.: Urban Institute.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*, 778-820.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2007). How and why do teacher credentials matter for student achievement? NBER Working paper #12828. Cambridge, MA: NBER.

Coates, D. (2003). Education production functions using instructional time as an input. *Education Economics*, *11*, 273-292.

Cohn. E. (1968). Economics of scale in Iowa high school operations. *Journal of Human Resources*, *3*, 422-434.

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*, *17*, 189-92.

Coleman, J.S., Campell, E.Q., Hobson, C.J., McParland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of Educational Opportunity*. U.S. Government Printing Office.

Creemers, B.P.M., & Scheerens, J. (1989). Developments in school effectiveness research. *International Journal of Education Research*, *13*, 685-825.

Cullen, J.B., Jacob, B.A., & Levitt, S.D. (2005). The impact of school choice on student outcomes: An analysis of Chicago public schools. *Journal of Public Economics*, *89*, 729-760.

DeCicca, P. (2007). Does full-day kindergarten matter? Evidence from the first two years of schooling. *Economics of Education Review*, *26*, 67-82.

Dee, T. (2004). Are there civic returns to education? *Journal of Public Economics*, *88*, 1697-1720.

Dreyfoos, J.G. (1990). *Adolescents at Risk: Prevalence and Prevention*. New York: Oxford.

Ehrenberg, R.G., & Brewer, D.J. (1994). Do school and teacher characteristics matter? Evidence from *High School and Beyond*. *Economics of Education Review, 13*, 1-17.

Ehrenberg, R.G., & Brewer, D. (1995). Did teachers' verbal ability and race matter in the 1960's? *Coleman* revisited. *Economics of Education Review*, *14,* 1-21.

Epple, D., & Romano, R.E. (1998). Competition between private and public schools, vouchers, and peer-group effects. *American Economic Review*, *88,* 33-62.

Evans, W.N., Oates, W.E., & Schwab, R.M. (1992). measuring peer-group effects: A study of teenage behavior. *Journal of Political Economy, 100*, 966-991.

Ferguson, R.F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal of Legislation*, *28*, 465-497.

Figlio, D. N. (2005). Boys named Sue: Disruptive dhildren and their peers. NBER working paper #11277. Cambridge, MA: NBER

Finn, J D. (1993). *School engagement and students at risk*. Washington, D.C.: National Center for Education Statistics.

Firestone, W.A. (1991). Introduction: Chapter 1. In J.R. Bliss, W.A. Firestone, & C.E. Richards (Eds), *Rethinking effective schools: Research and practice*. Englewood Cliffs, NJ: Prentice Hall.

Gardiner, M.F., Fox, A., Knowles, F., & Jeffrey, D. (1996). Learning improved by arts training. *Nature*, *381*, 254.

Garziano, A., Peterson, M., & Shaw, G.L. (1999). Enhanced learning of proportional math through music training and spatial-temporal training. *Neurological Research*, *21*, 139-152.

Goldhaber, D., & Brewer, D.J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, *32*, 505-523.

Gottfried, M.A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis*, *31*, 392-415.

Gottfried, M.A., & Inman, R.P. (2010). The effect of peers on student achievement in Philadelphia's classrooms. unpublished manuscript.

Greene, J.P. (1998). *A meta-Analysis of the effectiveness of bilingual education*. Los Angeles: Thomas Rivera Policy Institute.

Grogger, J. (1995). *School expenditures and post-schooling wages: evidence from High School and Beyond*. NSF Review of Economics and Statistics Conference Paper.

Grubb, W.N., & Flessa, J.J. (2006). 'A job too big for one': Multiple principals and other nontraditional approaches to school leadership. *Educational Administration Quarterly*, *42*, 518-550.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis*, *17*, 323-336.

Guttu, M., Engelke, M.K., & Swanson, M. (2004). Does the school nurse-to-student ratio make a difference? *Journal of School Health*, *74*, 6-9.

Haberman, M. (1991). The pedagogy of poverty versus good teaching. *Phi Delta Kappan*, *73*, 290-204.

Hallinger, P., & Heck, R. (1998). Exploring the principal's contribution to school effectivness: 1980-1995. *School Effectiveness and School Improvement*, *9*, 157-191.

Hanushek, E.A. (1972). *Education and Race: An Analysis of the Educational Production Process*. Health-Lexington: Cambridge, M.A.

Hanushek, E.A. (1979). Conceptual and empirical issues in the estimation of education production functions. *Journal of Human Resources*, *14*, 351-388.

Hanushek, E.A. (1986). The economics of schooling: production and efficiency in public schools. *Journal of Economics Literature*, *24*, 1141 – 1177.

Hanushek, E.A. (1989).The impact of differential expenditures on school performance. *Educational Researcher*, *18*, 45-62.

Hanushek, E.A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*, 84-117.

Hanushek, E.A. (1996). A more complete picture of school resource policies. *Review of Educational Research*, *66*, 397-409.

Hanushek, E.A. (2006). Teacher quality. In E.A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education*, *vol. 2* (pp. 2-28). Amsterdam: Elsevier North-Holland.

Hanushek, E.A., Kain, J.F., Markman, J.M, & Rivkin, S.G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18*, 527-544.

Hay/McBer (2000). *Research into teacher effectiveness*. Report for the Department of Education and Employment, United Kingdom.

Henderson, V., Mieszkowski, P., & Sauvageau, Y. (1978). Peer group effects and educational production functions. *Journal of Public Economics*, 10, 97-106.

Herman, B.E. (2004). The revival of K-8 schools. *Phi Delta Kappa Fastbacks*, *519*, 7-37.

Hoxby, C.M. (1996). How teachers' unions affect education production. *Quarterly Journal of Economics*, *111*, 671-718.

Hoxby, C.M. (2000). The effect of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics*, *115*, 1239-1285.

Jacob, B.A. (2007). The challenges of staffing urban schools with effective teachers. *Excellence in the Classroom*, *17*, 129-153.

Jencks, C., & Mayer, S.E. (1990). The social consequences of growing up in a poor neighborhood. In L.E. Lynn & M.G.H. McGeary (Eds.), *Inner-city Poverty in the United States*. Washington, D.C.: National Academy Press.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., & Michaelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York: Harper and Row.

Kearney, C.A., & Silverman, W.K. (1995). Family environment of youngsters with school refusal behavior: A synopsis with implications for assessment and treatment. *The American Journal of Family Therapy*, *23*, 59-72.

Lamb, S. (1997). Access to level of mathematics study in high school: Social area and school differences." In F. Biddulph & K. Carr (Eds.), *People in mathematics education*. Aotearoa, New Zealand: MERGA.

Lamb, S., & Fullarton, S. (2000). Classroom and teacher effects in mathematics achievement: Results from TIMSS. In J. Malone, J. Bana, & A. Chapman (Eds,), *Mathematics education beyond 2000*. Turramurra, Australia: MERGA.

Lamdin, Douglas J. (1996). Evidence of student attendance as an independent variable in education production functions. *The Journal of Educational Research, 89*, 155-162.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*, 37-62.

Lazear, E.P. (2001). Educational production. *Quarterly Journal of Economics*, *116*, 777-803.

Lee, J.W., & Barro, R.J. (1997). Schooling quality in a cross section of countries. NBER Working Paper #6198. Cambridge, M.A.: NBER.

Lee, V.E., & Smith, J.B. (1997). High school size: Which works best and for whom? *Educational Evaluation and Policy Analysis, 19*, 205-228.

Lehr, C.A., Hansen, A., Sinclair, M.F., & Christenson, S.L. (2003). Moving beyond dropout prevention to school completion: An integrative review of data based interventions. *School Psychology Review*, *32*, 342-364.

Lehr, C.A., Moreau, R.A., Lange, C.M., & Lanners, E.J. (2004). *Alternative schools: Findings from a national survey of the states.* Minneapolis, M.N.: University of Minnesota, Institute on Community Integration.

Lewis, S., Baker, N., & Jepson, J. (2000). *Critical trends in urban education: Fourth biennial survey of America's great city schools*. Washington D.C.: Council of Great City Schools.

Lewis, S., Baker, N., Jepson, J., Casserly, M., Powell, L., Barrengos, J., Johnson, J., & Eisner, C. (2000). *Great expectations: reforming urban high schools. An education forum with urban educators and leaders*. Washington D.C.: Office of Elementary and Secondary Education.

Link, C.R., & Mulligan, J.G. (1991). Classmates' effects on black student achievement in public school classrooms. *Economics of Education Review*, *10*, 297-310.

Lochner, L., & Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review*, *94*, 155-189.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economics and Statistics*, *60*, 531-542.

May, H., & Supovitz, J.A. (2006). Capturing the cumulative effects of school reform: An 11-year study on the impacts of America's choice on student achievement. *Educational Evaluation and Policy Analysis*, *28*, 231-257.

McEwan, P.J. (2003). Peer effects on student achievement: Evidence from Chile. *Economics of Education Review*, *22*, 131-141.

Moffitt, R.A. (2001). Policy interventions, low-level equilibria, and social interactions. In S. Durlauf, & H. Peyton Young (Eds.), *Social Dynamics*. MIT Press: Cambridge, M.A.

Monk, D.H. (1987). Assigning elementary pupils to their teachers. *Elementary School Journal*, *88*, 167-87.

Mortimore, P. (1991). The nature and findings of school effectiveness research in the primary sector. In S. Riddell & S. Brown (Eds.), *School Effectiveness Research: Research and Practice*. Englewood Cliffs, N.J.: Prentice, Hall

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years*. Wells: Open Books.

Nuttall, D.L., Goldstein, H., Prosser, R., Rabash, J. (1989). Differential school effectiveness. *International Journal of Educational Research, 13*, 769-776.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237-257.

Ogbu, J. U. (1989). The individual in collective adaptation: A framework for focusing on academic underperformance and dropping out among involuntary minorities. In L. Weis, E. Farrar, & H. G. Petrie (Eds.), *Dropouts from school: Issues, dilemmas, and solutions*. Albany, NY: State University of New York Press

Offenberg, R. (2001). The efficacy of Philadelphia's schools compared to middle grades schools. *Middle Schools Journal*, *37*, 14-23.

Orfield, G., Losen, D., Wald, J., & Swanson, C. B. (2004) *How minority youth are being left behind by the graduation rate crisis*. Washington, D.C.: The Urban Institute.

Peterson, B.B. (2002). School injury trends. *Journal of School Nursing*, *18*, 219-225.

Pew Charitable Trust (1998). The urban challenge: public education in the 50 states. *Education Week*, *17*, 1-270.

Phillips, L.M. (1996). Longitudinal effects of early literacy concepts on reading achievement: A kindergarten intervention and five-year follow up. *Journal of Literacy Research*, *28*, 173-195.

Reynolds, D., & Creemers, B. (1990). School effectiveness and school improvement: A mission statement. *School Effectiveness and School Improvement*, *1*, 1-3.

Riew, J. (1966). Economies of scale in high school operations. *Review of Economics and Statistics*, *43*, 280-287.

Rolstad, K., Mahoney, K., & Glass, G. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy, 19*, 572-594.

Rothstein, J. (2008). Student sorting and bias in value added estimation: Selection on observables and unobservables. unpublished manuscript.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith. A. (1979) *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge, M.A.: Harvard University Press

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics, 116,* 681-704.

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London, U.K.: Office for Standards in Education.

Scheerens, J. (1993). Effective schooling: Research, theory, and practice. *Educational Effectiveness and School Improvement*, *4*, 230-235.

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

Sheldon. S. B. (2007). Improving student attendance with school, family, and community partnerships. *The Journal of Education Research*, *100*, 267-275.

Smith, D.J., & Tomlinson, S. (1989). *The school effect*. London, U.K.: Policy Studies Institute.

Southhamer-Loeber, M., & Loeber, R. (1988). The use of prediction data in understanding delinquency. *Behavioral Sciences and the Law*, *6*, 333-354.

Summers, A.A., & Wolfe, B.L. (1977). Do schools make a difference? *American Economic Review*, *67*, 639-652.

Swanson, C. (2004). *Who graduates? Who doesn't? A statistical portrait of public high school graduation, class of 2001*. Washington, D.C.: The Urban Institute.

Tiebout, C.M. (1956). A pure theory of local expenditure. *Journal of Political Economy*, *64*, 416-424.

Tighe, E., Wang, A., & Foley, E. (2002). An analysis of the effect of Children Achieving on student achievement in Philadelphia elementary schools. Philadelphia, PA: Consortium for Policy Research in Education.

Tobin, K., Seiler, G., & Walls, E. (1999). Reproduction of social class in the teaching and learning of science in urban high schools. *Research in Science Education*, *29*, 171-187.

Todd, P.E., & Wolpin, K.I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *113*, F3-F33.

Waxman, H.C., & Padron, Y.N. (1995). Improving the quality of classroom instruction for students at risk of failure in urban schools. *Peabody Journal of Education, 70,* 44-65.

Woβmann, L. (2003). Schooling resources, educational institutions, and student performance: the international evidence, *Oxford Bulletin of Economics and Statistics*, *65*, 117-170.

Zimmer, R. (2003). A new twist in the educational tracking debate. *Economics of Education Review*, *22*, 307-315.

Zimmerman, D. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *The Review of Economics and Statistics*, *85*, 9-23.