Peer Effects in Disadvantaged Primary Schools: Evidence from a Randomized Experiment

Heather Antecol*
Claremont McKenna College and IZA

Ozkan Eren
Louisiana State University

Serkan Ozbeklik
Claremont McKenna College

December 5, 2013

Abstract

We use data from a well-executed randomized experiment in seven states to examine the effect of peer achievement on students' own achievement in primary schools in disadvantaged neighborhoods. Contrary to the existing literature, we find that the average classroom peer achievement *adversely* influences own student achievement in math and reading. Extending our analysis to take into account the potential nonlinearity in the peer effects leads to non-negligible differences along the achievement distribution. We test several models of peer effects to further our understanding of peer interactions in primary schools. While we find no evidence to support the monotonicity model and little evidence in favor of the ability grouping model, we find stronger evidence to support the frame of reference and the invidious comparison models. Finally, using a unique feature of our data, we attempt to reconcile our findings with the existing literature.

JEL Codes: I21, J24
Key Words: Peer Effects; Student Achievement; Random Assignment; Disadvantaged Schools

*Corresponding Author. All future correspondence regarding this manuscript may be addressed to Heather Antecol, Boswell Professor of Economics, The Robert Day School of Economics and Finance, Claremont McKenna College, 500 E. Ninth St., Claremont, CA 91711. Tel: (909) 607-8067. Fax: (909) 621-8249. E-mail: hantecol@cmc.edu.

1. Introduction

Throughout the past several decades there has been a nation-wide debate focusing on how to improve student achievement in the United States. The debate was fueled by the influential Coleman Report of 1966 which questioned the long-standing belief that school funding was a key determinant of student achievement.[1] The report instead highlighted the importance of alternative determinants—e.g., family background and socio-economic status, teacher quality, and peer quality—which could have differential effects on students in schools in economically disadvantaged neighborhoods relative to students in schools in more economically advantaged neighborhoods (Coleman 1966). Not surprisingly, the report spawned a flurry of new research among social scientists, as well as a shift in policy-makers' education goals. However, it continues to be the case that there is little, if any, agreement about which specific education policies are more effective in improving student achievement (for reviews of the literature see Hanushek 2006; Hanushek and Rivkin 2006). One set of education policies that have received substantial attention are those that result in a change in the mix of one's peers (e.g., ability tracking, school choice programs, and racial and economic integration). Intuitively, it is unclear if this change in peers will impact student achievement for all students equally or if it will have differential effects on student achievement depending on a student's own achievement and background (e.g., gender, race, socio-economic status). This uncertainty highlights the importance of understanding the influence of peers on student achievement.

The estimation of the causal effect of peers however is plagued with difficulties. In particular, any study attempting to measure the causal effect of peer quality on student achievement has to deal with two important identification issues. First, it is a well-known fact

[1] This debate has gained further momentum in the last decade as American students' test score outcomes continue to lag behind their counterparts from many other developed countries (Fleischman et al. 2010)

that students are not randomly assigned to schools or classrooms largely because of families, school administrators, or teachers (see for example, Clotfelter, Ladd, and Vigdor 2006; and Kane et al. 2011). This identification issue is often referred to as the selection problem (Sacerdote 2001). Failure to account for non-random sorting of students in a regression framework would result in biased coefficient estimates of peer effects as there are likely to be observable and unobservable factors that affect both student achievement and peer quality. Second, it is often difficult to disentangle the impact that the peer group has on the student from the impact the student has on the peer group. A regression of, say, own achievement on contemporaneous average achievement of peers is problematic as these outcomes are jointly determined and peer achievement is likely to be endogenous in the model. This is usually referred to as the endogeneity or the reflection problem (Manski 1993; Moffitt 2001; Sacerdote 2001).

The existing literature on peer effects and student outcomes in grades K through 12 (henceforth referred to as the existing literature) generally relies on panel or repeated cross-sectional data sets and uses within-school/grade variation in achievement or some other school/grade characteristics to measure peer effects to overcome the threats to identification (see for example, Hanushek et al. 2003; Vigdor and Nechyba 2007; Lavy, Silva, and Weinhardt 2012; Lavy, Paserman, and Schlosser 2012; and Burke and Sass 2013). To the extent that the within-school or grade variation is random, the coefficient estimates on peer measures produce reliable estimates of peer effects. There are also a handful of studies trying to solve the identification problems by exploiting natural experiments and/or random assignment of students to groups (see for example, Hoxby and Weingarth 2006; Duflo, Dupas, and Kremer 2011; Imberman, Kugler, and Sacerdote 2012; and Jackson 2013). Moreover, methodological limitations and data constraints compel many studies to focus on only one state and/or school district.[2] Finally, with the exception of Hoxby and Weingarth (2006), Duflo, Dupas, and Kremer (2011), Imberman, Kugler, and Sacerdote (2012), and Lavy, Paserman, and Schlosser (2012), the existing literature has primarily focused on documenting the existence of peer effects as opposed to formally identifying the underlying models of peer effects.

The findings in the existing literature are mixed at best. While some studies find positive and significant effects of average peer achievement on students' own achievement (see for example, Hoxby 2000; Boozer and Cacciola 2001; Hanushek et al. 2003; Betts and Zau 2004; Hoxby and Weingarth 2006; Vigdor and Nechyba 2007; Carman and Zhang 2012; and Lavy, Paserman, and Schlosser 2012), others find small to no effects (see for example, Lavy, Silva, and Weinhardt 2012; Imberman, Kugler, Sacerdote 2012; and Burke and Sass 2013). The common perception from several of these studies is that it is not only the high ability students but also those at the bottom of the achievement distribution who seem to benefit from higher achieving peers. With that said, the peer effects estimates are not identical across different achievement groups and the impacts generally exhibit nonlinearities with no consensus on who benefits the most from better peers (see for example, Imberman, Kugler, and Sacerdote 2012; Lavy, Silva, and Weinhardt 2012; Jackson 2013; and Burke and Sass 2013).[3,4]

[2] For example, Burke and Sass (2013) examine public school students in grades 3-10 in Florida; Hoxby and Weingarth (2005) examine the Wake County Public School district in North Carolina;; Betts and Zau (2004) examine the San Diego Unified School District in California. In contrast, Imberman, Kugler and Sacerdote (2012) examine primary schools in two states: the Houston Independent School District in Texas and Louisiana.

[3] There is also a large literature examining the effect of peers on student outcomes in college. The results from these studies are again mixed. Studies either find small positive effects (Sacerdote 2001; Zimmerman 2003), large positive effects (Stinebrickner and Stinebrickner 2006; Carrell, Fullerton, and West 2009), or no effects (Foster 2006; Lyle 2007). Moreover, there are a number of recent studies that examine peer effects in labor markets (see for example, Arcidiacono and Nicholson 2005; Bandiera, Barankay and Rasul 2005; Falk and Ichino 2006; Mass and Moretti 2009; Guryan, Kroft, and Notowidigdo 2009; and Brown 2011) and on social and behavioral outcomes (see for example Case and Katz 1991; Gaviria and Raphael 2001; Ludwig, Duncan and Hirsfield 2001; and Kling, Ludwig, and Katz 2005).

[4] For a detailed review of the empirical peer effects literature see Sacerdote (2011) and of the theoretical peer effects literature see Epple and Romano (2011).

The purpose of this paper is to contribute to the existing literature on the effect of peer achievement on students' own achievement in primary schools in the following ways. First, this is the first study to the best of our knowledge that explicitly examines peer effects and achievement in primary schools in *disadvantaged* neighborhoods, as well as focuses on more than one state/school district. Our focus on these disadvantaged neighborhoods is deliberate. Specifically, it is well-documented that students in disadvantaged neighborhoods have lower achievement levels relative to their affluent counterparts (see for example, Hanushek and Raymond 2005, Curto and Fryer 2013; and Sass et al. 2012). As such, this is the segment of the population that is in the need of the most help and has been the target of many policy initiatives including the Obama Administration's Race to the Top Program. Therefore focusing on disadvantaged neighborhoods allows us to take a closer look at the influence of peers on student achievement in a setting where the problems with the education system in the United States are most evident and arguably more important from a policy perspective. Moreover, many studies highlight the fact that the returns to inputs in the educational production function (such as, parental involvement, teacher qualifications and class size) vary considerably for children from disadvantaged neighborhoods compared to those from affluent neighborhoods (see for example, Krueger and Whitmore 2001, Clotfelter, Ladd and Vigdor 2006 and Sass et al. 2012). Given this, there is no a priori reason to believe that peer interactions—which are another input in the education production function—will operate the same way in environments that differ economically.

Second, our data comes from a well-executed randomized experiment which helps us avoid the aforementioned selection problem, and thus allows us to measure the causal effect of peer quality on student achievement. Third, unlike many existing studies, we measure peer achievement at the classroom level which is arguably a better approximation of the peer interactions in primary schools than the grade level peer achievement measure traditionally employed.[5] In particular, children spend at least six hours a day for roughly 180 days a year with their classmates while the time they interact with their other schoolmates is rather limited and usually only occurs during the recess. Finally, following Hoxby and Weingarth (2006) and Imberman, Kugler, and Sacerdote (2012), we also explicitly examine how peer effects might work in disadvantaged neighborhoods. Specifically, we focus on four potential models of peer effects: the monotonicity model (i.e., the effects of peers on student achievement is increasing in peer quality); the invidious comparison model (i.e., higher ability peers adversely influence the outcomes of students who are moved to a lower position in the local achievement distribution), the ability grouping (boutique) model (i.e., student performance is highest when their peers are similar to themselves), and the frame of reference model or the reverse big fish in a little pond model (i.e., higher ability peers adversely influence the outcomes of students due to a lower academic self-concept).

Contrary to the existing literature, we find that the average classroom peer achievement *adversely* influences own student achievement irrespective of subject or group, although the effect is imprecisely estimated for certain subgroups. Extending our analysis to take into account the potential nonlinearity in the peer effects leads to non-negligible differences along the achievement distribution. Focusing first on reading test scores, we find that an improvement in peer quality for the full sample substantially hurts students both at the bottom and top of the achievement distribution but does not seem to affect middle ability students. The subgroup patterns for students in all achievement groups essentially mirror those found for the full sample,

[5] Most of the studies in the peers effect literature prefer to use grade level peer measures as a further breakdown of the peer interactions to say classrooms requires one to control for the potential nonrandom sorting of students to classrooms. Notable exceptions are Hoxby and Weingarth (2006) and Burke and Sass (2013).

however the effects are estimated more precisely for certain subgroups (particularly at the top of the distribution). Turning to math test scores, we find negative effects of peer quality for the full sample over the entire achievement distribution, although the coefficient estimates are imprecisely estimated. The patterns for the full sample appear to extend only to male students at all achievement levels and Hispanic students for the middle achievement group for whom we observe negative (and significant) peer effects. Viewing the complete set of results, we find no evidence to support the monotonicity model and little evidence in favor of the ability grouping model, while we find stronger evidence to support the frame of reference and the invidious comparison models. Finally, in an attempt to reconcile our results with those in the previous literature, we use a unique feature of our data and exploit cohort-to-cohort variation in peer achievement within the same school and grade to identify the peer effects, as well as allow for differences in the level of peer aggregation. While we find that the importance of identifying the salient peer group cannot be understated in estimating peer effects, our overall findings from these exercises suggest that neither the potentially confounding effects in repeated cross-section data nor the use of a classroom level peer measure (as opposed to a grade level measure) appear to be valid explanations. These exercises do however provide tentative evidence that our focus on students in primary schools in *disadvantaged* neighborhoods may potentially be the driving force behind the divergence in our results and the results in the existing literature.

2. Data and Tests for Random Assignment

2.1 Data

We use data from the Mathematic Policy Research, Inc. (MPR) Teachers Trained Through Different Routes to Certification (TTTDR) Private Use File. TTTDR is a randomized study of primary school students, which was conducted to assess the effectiveness of different teacher certifications on student outcomes. MPR began in 2003 by identifying as many schools with alternatively certified (AC) teachers as possible where AC teachers are those who become a classroom teacher prior to completing all required coursework and without having to complete a period of student teaching.[6] In order to be eligible for the study, (i) schools had to have had at least one alternative certification (AC) and one traditional certification (TC) teacher in the same grade (i.e. kindergarten through grade 5); (ii) both AC and TC teachers had to have had five or fewer years of experience, and (iii) both AC and TC teacher must have taught in regular classes and must have delivered both math and reading instruction to all their own students.[7]

MPR identified 170 schools meeting the eligibility criteria for the study. Among this compiled list of eligible schools, a stratified random sample of 60 schools was selected. Specifically, in the spring of 2004 the study administrators contacted schools to search for a suitable pair of teachers who could potentially be in the study for the 2004-2005 school year and these efforts yielded a sample of 20 AC and 20 TC teachers in 20 schools. For the 2005-2006 school year, MPR retained as many teachers as possible from the first year (10 teachers) and recruited additional teachers from the same school (10 schools total), as well as from schools in the same school district and from new school districts. It is important to note that retained teachers teach the same grade in both the first year and the second year but they have new classrooms with randomly assigned students. The final sample included 90 AC and 90 TC

[6] The AC programs differ on the selectivity criteria of their admission requirements. For instance, AC programs such as the Teach for America require a minimum GPA of 3.0 from the applicants. The AC teachers in the TTTDR sample come from programs with less selective entrance requirements by design as this maintained a fairer comparison between AC and TC teachers. We further note that the TTTDR study did not find any difference in the end of the academic year test scores between students taught by AC and TC teachers.

[7] Even though the requirements for teachers who pursue alternative routes to certification vary by state and district, the AC programs, on average, require significantly less education coursework than TC programs (see Constantine et al. 2009 for more details on AC and TC teachers).

teachers and more than 2,800 students that were selected in seven states between 2004 and 2006.[8,9]

This data is ideal for our purposes because, within each school, all students in the same grade were randomly assigned to either an AC or a TC teacher before the start of the academic year. Therefore, the randomization is done at the block level such that each block represents classrooms in the same grade level in any given school. This process not only ensured that those students in AC and TC classrooms are comparable but also that the pre-treatment achievement of own students and the average pre-treatment achievement of their peers in each classroom are not correlated (this is discussed in further detail in Section 2.2). We have a total of 90 blocks of which 94 percent are pairs (1 AC and 1 TC classroom), 4 percent are trios (3 classrooms with at least one being an AC and one being a TC classroom), and 2 percent are quartets (2 AC and 2 TC classrooms).

After the random assignment and before the start of the academic year, the students were given math and reading tests based on the grade they completed in the previous year (which we call *baseline* outcome variables); then at the end of the academic year in which the study was conducted the students re-took math and reading tests based on the grade they just completed (which we call endline outcome variables). We use Normal Curve Equivalent (NCE) points in math and reading as our measures of baseline and endline test scores.[10] The NCE scale has a mean of 50 and standard deviation of 21 nationally.

[8] Due to the confidential nature of the data agreement, the sample sizes are rounded to the nearest tenth.

[9] The states included in the TTTDR sample are California, Georgia, Illinois, Louisiana, New Jersey, Texas and Wisconsin. There were 20 school districts in the effective sample; 5 districts from California, 7 districts in total together from Georgia, Illinois, Louisiana and Wisconsin, 3 districts from New Jersey and 5 districts from Texas.

[10] The students were administered two reading tests (reading comprehension and vocabulary). The sum of scores from these two tests establishes total reading score and our measure of student achievement in reading. There were also two different tests in math (math concepts and applications and math computation). Unlike reading,

The sample attrition in TTTDR data set is relatively small, but we still lose roughly 7 (8) percent of the initial reading (math) sample because of missing test scores.[11] After dropping these observations, our estimation sample consists of 2,610 (2,580) students for the reading (math) test score sample from classes taught by 180 teachers. To ensure the student composition was unaffected by the sample attrition, Constantine et al. (2009) show the attrition rates in the AC and TC samples were almost identical and did not differ significantly between the two types of classrooms (Appendix A, pp. A13, Table A3 in Constantine et al. 2009).[12]

Besides test scores and the type of classroom (AC or TC classroom) the data set also contains information on the student's gender, race/ethnicity, and eligibility for free lunch. Table 1 present some features of the TTTDR student sample. Specifically, 34.5 (47.0) percent of the student body is black (Hispanic), while 9.2 percent is white.[13] Moreover, students tend to come from low income families; roughly 75 percent of the effective sample is eligible for free lunch as opposed to 40 percent nationwide. Finally, the average baseline test scores for reading and math are roughly 39 and 42 NCE points for the full sample, respectively. Compared to the national average, the reading (math) scores are roughly 0.5 (0.4) of a standard deviation lower in the TTTDR sample. Overall, it is evident that the TTTDR sample consists of lower achieving students from disadvantaged neighborhoods.

The second and third columns of Table 1 report the average baseline characteristics of students in AC and TC classrooms, respectively. Under the assumption that the random assignment is implemented correctly, baseline characteristics of students in AC and TC

however, students in kindergarten and grade 1 were not administered math computation test. Thus our measure in math achievement is scores from math concepts and applications only.

[11] Students test scores are missing either because they moved out of school district or they did not take endline tests.

[12]Ideally, we would like to run a regression of the non-response indicator on an AC classroom dummy along with the baseline characteristics. Even the restricted version of the data set, however, does not include any information on those moving out of the school district and on students not taking the test.

[13] The remaining 9.3 percent of the student body indicated "other" race. The survey instrument does not provide details on what this category includes.

classrooms must be similar. To test this, as in Krueger and Whitmore (2001), we run a regression of the AC indicator variable on each baseline characteristic conditional on block fixed effects (the dependent variable taking the value of one if the student is in a AC classroom and zero otherwise). The fourth column of Table 1 displays the coefficient estimates from this exercise. None of the coefficient estimates are statistically significant at conventional levels. By the nature of randomization in TTTDR, it is important to note that we include the block fixed effects in all of our specifications throughout the paper. The use of conditional randomization is a very common practice in the education literature (see for example, Sacerdote 2001; Carrell, Fullerton, and West 2009; and Duflo, Dupas and Kremer 2011).

Finally, TTTDR includes information on teacher characteristics including gender, race/ethnicity, teaching experience, hours of instruction for certification, and SAT Composite Score. Not surprisingly given our sample is comprised of primary schools, roughly 90 percent of teachers are female (see Column 1 of Appendix Table A1), however AC teachers are less likely to be female relative to their TC counterparts (see Columns 2 and 3 of Appendix Table A1). TC teachers are less racially/ethnically diverse than their AC counterparts. Specifically, roughly 72 (45) percent of TC (AC) teachers are white. By construction AC and TC teachers have similar levels of teaching experience, roughly 3 years. Finally, TC teachers have roughly 2 times more teaching training than their AC counterparts, although this difference is somewhat less pronounced for math.

2.2 Are Peers Randomly Assigned?

Although we have shown some preliminary evidence on the random assignment of students within two types of classrooms, it is imperative for the purpose of our study to validate the random assignment of peers (absence of sorting) within blocks. A typical test for this is to run an OLS regression of student *i*'s pre-determined achievement on the pre-determined average achievement of *i*'s peers, controlling for any variable on which randomization was conditioned on and is given by

$$TS_{icb}^{base} = \pi_0 + \pi_1 \overline{TS}_{-i,cb}^{base} + \eta_b + u_{icb} \qquad (1)$$

where TS_{icb}^{base} is the subject-specific baseline test score for student *i* in classroom *c* and block *b*, $\overline{TS}_{-i,cb}^{base}$ is the average peer baseline subject-specific test score in classroom *c* and block b excluding student *i*, η_b is a set of block fixed effects (i.e., classrooms in the same grade level in any given school), and u_{icb} is the error term. Under the assumption that peers are randomly assigned, one would expect the estimate of π_1 to be equal to zero. This approach is the common practice in the peer effects literature to test for randomization (see for example, Sacerdote 2001; Foster 2006; and Carrell, Fullerton, and West 2009).

Guryan, Kroft, and Notowidigdo (2009) however recently showed that the mechanical relationship between own ability and average ability of peers (i.e., peers of high achieving students are chosen from a block with a slightly lower mean achievement than peers of low achieving students) may cause the aforementioned falsification exercise to produce negative and statistically significant coefficient estimates for π_1. Random assignment may not appear random, while positive sorting of students to classrooms may appear random.[14]

Given the bias is a by-product of the differences in the average achievement level of the group once the student *i* is withdrawn, the proposed solution in Guryan, Kroft, and Notowidigdo

[14] It is also important to note that as the size of the randomization group (block in our case) grows the contribution of each student to the average ability goes down and the magnitude of the bias from the falsification exercise is also reduced. The average classroom and block sizes in our study are 15.1 and 32, respectively.

(2009) is to control for this relevant group mean in the falsification regressions. Specifically, the revised falsification test equation is given by

$$TS_{icb}^{base} = \pi_0 + \pi_1 \overline{TS}_{-i,cb}^{base} + \pi_2 \overline{TS}_{-i,b}^{base} + \eta_b + u_{icb} \qquad (2)$$

where $\overline{TS}_{-i,b}^{base}$ is the mean achievement of students in block b and all other variables are as previously defined. Using simulations, Guryan, Kroft, and Notowidigdo (2009) show that equation (2) is a well-behaved randomization test and if the student assignment to classrooms is truly random, we would expect the coefficient estimate $\hat{\pi}_1$ to be equal to zero.

Table 2 presents our results from various falsification tests. The first and second columns of Table 2 report the results from estimating equation (1) for baseline reading and math test scores, respectively, while the third and fourth columns report the results from estimating equation (2). In the absence of correction, the correlation between own and peers' baseline achievement is negative and statistically significant for both the reading and math test score samples. As previously noted, however, ignoring the bias in the randomization tests leads to the erroneous conclusion that students are negatively sorted within each block. Once we do the correction the coefficient estimates on average peer baseline test scores are insignificant and almost equal to zero in magnitude irrespective of subject.

To examine the integrity of the experiment, we also replace the achievement measures in equations (1) and (2) with several selected peer background characteristics (i.e., share of female students in the classroom; share of Hispanic students in the classroom; share of black students in the classroom, and the share of free lunch eligible students in the classroom). Appendix Table A2 present these additional tests and none of the coefficient estimates in these specifications are different than zero. This provides further evidence on the random assignment of peers to classrooms (see for example, Carrell and Hoekstra 2010 for a similar falsification exercise). The coefficients from all of the falsification tests also remain intact when we control for classroom type (AC or TC classroom).

3. Empirical Methodology and Results

3.1 Empirical Methodology

Having shown that students are randomly assigned to classrooms, we now turn to the estimation of peer effects on student achievement. To begin with, we first analyze the peer effects using linear-in-means models, where we regress endline test scores on average peer baseline test scores along with students' own baseline scores and block fixed effects. In a randomized experiment setting, it is a well-known fact that controlling for the baseline characteristics does not affect the consistency of the estimates; however, it helps increase efficiency (Frölich and Melly, 2013). To this end, we estimate the following equation for the full sample and by subgroups (i.e., student gender, student race/ethnicity, and student free lunch eligibility status):

$$TS_{icb}^{end} = \beta_0 + \beta_1 \overline{TS}_{-i,cb}^{base} + \beta_2 TS_{icb}^{base} + SC_{icb}^{'}\delta + TC_{cb}^{'}\gamma + \eta_b + e_{icb} \qquad (3)$$

where TS_{icb}^{end} is the subject-specific endline test score for student *i* in classroom *c* and block *b*. *SC* is a set of student characteristics (i.e., gender, race/ethnicity, and free lunch status), *TC* is a set of teacher characteristics (i.e., AC/TC status, gender, race/ethnicity and years of teaching experience), TS_{icb}^{base} , $\overline{TS}_{-i,cb}^{base}$, and η_b are as previously defined. It is important to note that peer effects estimates are reduced form in the sense that equation (3) does not separately identify the effects of peer outcomes (endogenous effects) and peers' background characteristics (contextual effects).

We also estimate two versions of equation (3) for the full sample and by subgroups to address the potential nonlinearity in the peer effects. The first version is given by

$$E(TS_{icb}^{end} \mid Q_k^{base}) = \beta_0 + \beta_1 \overline{TS}_{-i,cb}^{base} + \beta_2 TS_{icb}^{base} + SC_{icb}^{'}\delta + TC_{cb}^{'}\gamma + \eta_b + e_{icb} \tag{4}$$

where Q_k^{base} is the student *i*'s grade and subject-specific baseline achievement quartile *k* (*k* = top 25%; middle 25-75%; bottom 25%) and all remaining variables are as previously defined. We estimate equation (4) separately for each quartile. The second version specifies a slightly different measure of peer quality and the estimation equation is given by

where

$$E(TS_{icb}^{end} \mid Q_k^{base}) = \beta_0 + \beta_{k,bottom} P_{-i,cb}^{bottom} + \beta_{k,top} P_{-i,cb}^{top} + \beta_2 TS_{icb}^{base} + SC_{icb}^{'}\delta + TC_{cb}^{'}\gamma + \eta_b + e_{icb} \tag{5}$$

$P_{-i,cb}^{bottom}$ and $P_{-i,cb}^{top}$ represent the fraction of bottom 25% and top 25% of peers in classroom *c* and block b, respectively, based on the grade and subject-specific baseline test score distribution. Due to collinearity of the proportions, we omit the proportions of middle ability peers in each specification of equation (5). All other variables are defined as previously. Finally, we report the standard errors clustered at the block-level beneath each coefficient estimate.

We examine four potential models through which peer effects might work. First, we examine the monotonicity model which implies that the effect of peers on student achievement is increasing in peer quality. Using equation (5) we test for two versions of the monotonicty model: weak monotonicity states $\beta_{k;top} > \beta_{k;bottom}$ *and* strong monotonicity states $\beta_{k;top} > \beta_{k;middle}$ and $\beta_{k;middle} > \beta_{k;bottom}$. for k = *top; middle; bottom.*

The second model we examine is the invidious comparison (proposed in Hoxby and Weingarth 2006) which states that higher ability peers adversely influence the outcomes of students who are moved to a lower position in the local achievement distribution, perhaps because of a fall in their self-esteem. Note that this model does not say anything about the impact of peers at the same ability level. Using equation (5) we test for the invidious comparison model as follows: for k = *top; middle; bottom* and j = *top; middle; bottom*, the invidious comparison model states $\beta_{kj} < 0$ for $j > k$ and $\beta_{kj} > 0$ for $k > j$ where j denotes the grade and subject-specific baseline achievement quartile of peers.

The third model we examine is the ability grouping (boutique) (proposed in Hoxby and Weingarth 2006) which states that student performance is highest when their peers are similar to themselves. Using equation (5) we test for the ability grouping model as follows: for k = *top; middle; bottom* and j = *top; middle; bottom*, the ability grouping model states $\beta_{kk} > \beta_{kj}$ for $j \neq k$.

The final model we examine relies on social comparison theory and frame of reference (Marsh and Parker 1984; Marsh 1987). In an educational setting, the theoretical model underlying the frame of reference model states that students compare their own academic achievement with the achievement of peers and use this social comparison for forming their own academic self-concept where academic self-concept is defined as one's knowledge and perceptions about one's academic ability. In this context, academic self-concept depends not only on one's own achievement but also on the achievement of a reference group. Consider a high achieving student in a regular classroom is assigned to a gifted classroom; the student in this new environment may become an average student relative to his peers. According to Marsh and Hau (2003), this then can have adverse effects on the student's academic self-concept as he is no longer “a big fish in a small pond” (regular class) but is now “a little fish in a big pond” (gifted class). According to the frame of reference model, academic self-concept will be affected positively with individual achievement but will also be negatively affected by the average achievement of the reference group. Thus the frame of reference model predicts a negative impact of an improvement in peers’ achievement on student's own achievement.

Taking this a step further, if the proportion of peers in the top (bottom) 25% increases, then average peer achievement must improve (decline) which will result in a negative (positive) impact on own student achievement. In other words, the frame of reference model predicts that all students are hurt from high-achieving peers and benefit by low-achieving peers (i.e., the inverse of the monotonicity model).We test models of both the weak and strong frame of reference as follows: for *k* = *top; middle; bottom*, the weak frame of reference states $\beta_{k;top} < \beta_{k;bottom}$ and for *k* = *top; middle; bottom*, the strong frame of reference states $\beta_{k;top} < \beta_{k;middle}$ and $\beta_{k;middle} < \beta_{k;bottom}$.[15]

3.2 Results

3.2.1 Linear-in-Means Results

Column 1 of Panel A and B of Table 3 present our linear-in-means estimations for reading test scores and math test scores, respectively, for the full sample. Specifically, the coefficient estimate on average classroom peer baseline reading achievement is negative and statistically significant (-0.18); a one standard deviation increase in peer achievement is associated with roughly one-ninth of a standard deviation decrease in own endline reading scores. Similarly, the coefficient estimate on average classroom peer baseline math achievement is (-0.24) suggesting that a one standard deviation increase in peer achievement decreases math test scores by around one-ninth of a standard deviation as well. For both reading and math test scores, we find very similar results if we exclude teacher characteristics only or if we exclude

[15] Hoxby and Weingarth (2006) outline a number of other potential models including the bad apple model (i.e., one disruptive student has a detrimental effect on the outcomes of all students irrespective of where they are in the achievement distribution); the shining light model (i.e., one excellent student has a positive effect on the outcomes of all students irrespective of where they are in the achievement distribution); the focus model (i.e., homogeneous classrooms are good irrespective of student *i*'s ability relative to their homogeneous peers); and the rainbow model (i.e., heterogeneous classrooms benefit all students).

both student and teacher characteristics (see Columns 1 through 3 of Appendix Table A3). This provides further evidence that the assignment is truly random at the block level as the additional controls simply add precision to the model. Moreover, we also estimate a specification that adds average block baseline peer achievement in addition to average classroom peer baseline achievement (see Column 4 of Appendix Table A3). The coefficient estimate on average block peer achievement is not different than zero and the average classroom peer baseline achievement remains virtually unchanged.

To further explore these findings, we extend our analysis to test for the presence of heterogeneous effects along a number of dimensions. We first focus on student gender. While we continue to find a negative effect of average classroom peer achievement, irrespective of subject or student gender, the magnitude of the peer effect is substantially larger for male students and is imprecisely estimated for female students (see Column 1 of Tables 4 and 5 for female and male students, respectively). These gender differences may stem from the fact that female students tend to be more cooperative and more level even in the presence of ability differences with their peers (see for example, Croson and Gneezy 2009 and Bertrand 2010).

Our next set of results pertains to student free-lunch status, which proxies for family income. We noted earlier that roughly 75 percent of the effective sample is eligible for free-lunch and students from wealthier families are disproportionately distributed at the top of the achievement group. This leaves us with only a limited number of observations at the bottom quartile for students that are not eligible for free-lunch which is particularly important when we allow for nonlinearities. As such, we focus our discussion on free-lunch eligible students only. We again find that, irrespective of subject, that average classroom baseline peer achievement adversely influences own endline test scores (see Column 1 of Tables 6).

Our final set of results pertains to student race/ethnicity. As previously noted only 9 percent of our estimation sample comprises white students. Such a small sample prevents us from making a rigorous inference and therefore we only focus on black and Hispanic students. For both black and Hispanic students, we continue to find a negative effect of average classroom peer achievement for reading and math test scores, although the effects are imprecisely estimated (see Column 1 of Tables 7 and 8 for black and Hispanic students, respectively).

3.2.2 Nonlinearities in Peer Effects

In the previous section, we assume the peer effects are linear. There is, however, substantial evidence against the linear-in-means model (see for example, Hoxby and Weingarth 2006; Sacerdote 2011; and Imberman, Kugler, and Sacerdote 2012). Moreover, from a policy point of view, if the peer effects were to be linear, there would be no gain or loss in sorting and tracking students. To examine the potential nonlinearities in peer effects, we first estimate the effect of average classroom subject specific baseline peer achievement on own subject specific endline test scores separately by the grade and subject-specific baseline achievement quartile k (k = top 25%; middle 25-75%; bottom 25%) for student i (see equation 4 in Section 4.1).[16]

Focusing first on the full sample estimation, we observe a negative and significant impact of average classroom peer baseline reading achievement for students at the bottom quartile of the achievement distribution (-0.45); a one standard deviation increase in peer achievement is associated with roughly one-fourth of a standard deviation decrease in own reading test scores (see Column 3, Panel A of Table 3). The coefficient estimate on peer effects for the middle achievement group, on the other hand, is almost equal to zero in magnitude (-0.06) and is

[16] We discuss alternative cut-off points to describe the bottom and top achievement groups in Section 3.2.5. However, we are unable to examine cut-off points based on the top 5 (10) % and bottom 5 (10) % due to data limitations (i.e., our sample size does not allow us to cut the data that finely).

insignificant (see Column 5, Panel A of Table 3), while the effect for the students in the top quartile is negative (-0.26) although imprecisely estimated at conventional levels (see Column 7, Panel A of Table 3). Pair-wise comparisons indicate that the peer effects coefficient for the lowest achievement group is significantly different than the one for middle achievement group (p-value 0.02). Turning to the math test score results (see Column 3, 5, and 7 of Panel B of Table 3), the coefficient estimates are negative and similar in magnitude for all achievement groups, although imprecisely estimated at conventional levels. We fail to reject the null of equality across all pair-wise comparisons of peer effects coefficient estimates.

The patterns for the full sample generally extend to all subgroups under consideration. As such, we only highlight what we believe to be the most interesting findings here (see Columns 3, 5, and 7 of Tables 4-8). For male students, the peer effects for reading is large and more precisely estimated at the top of the achievement distribution (see Column 7, Panel A of Table 5). Similarly, the peer effect for math is larger in all three quartiles of the distribution, although the effect continues to be imprecisely estimated in the top quartile (see Columns 3, 5, and 7, Panel B of Table 5). For free-lunch eligible students, an improvement in classroom reading achievement hurts the students at the top quartile the most (see Columns 3, 5, and 7, Panel A of Table 6). Moreover, it appears that an increase in peer math quality decreases the achievement level of Hispanic students at the middle quartile (see Column 5, Panel A of Table 8). Finally, we observe a positive coefficient estimate for Hispanic students who are at the top quartile of the math score distribution but this coefficient estimate is imprecisely estimated (see Column 7, Panel A of Table 8).

To further delve into the complexity of the effect of peers on student achievement we replace average classroom baseline achievement with the fraction of bottom 25% and top 25% of peers in classroom *c*, respectively, based on the grade and subject-specific pre-treatment test score distribution. Due to collinearity, we omit the proportion of middle ability peers in each regression (see equation 5 in Section 4.1). We first replace average peer achievement with these fractions irrespective of student *i*'s placement in the grade-subject specific baseline achievement distribution (Column 2 of Tables 3-8). We then also allow student *i*'s placement in the grade-specific baseline achievement distribution to vary as well (Columns 4, 6, and 8 of Tables 3-8). We focus on the results for the full sample and discuss any significant deviations from these results for the subgroups under consideration in the remainder of this section.

If we hold student *i*'s placement in baseline grade and subject-specific achievement distribution fixed (see Column 2 Table 3), we find that a 1 percentage point increase in the proportion of peers in the top quartile (therefore the proportion of peers in the middle quartile decreases by 1 percentage point) is associated with a 0.048 (0.018) points decrease (increase) in endline reading (math) test scores, although the effects are imprecisely estimated.[17] Similarly, if the proportion of peers in the bottom quartile increases by 1 percentage point relative to the middle quartile, then endline reading (math) test scores go up by 0.047 (0.115) points, the effect however is statistically insignificant at conventional levels for reading test scores.

If we now also allow student *i*'s placement in the baseline grade and subject specific achievement distribution to vary, for students in the bottom quartile we find that if you increase the proportion of peers in the top (bottom) quartile by 1 percentage point relative to the middle quartile, then endline reading tests scores decrease (increase) by 0.06 (0.14) points (see Column 4, Panel A of Table 3), the effect however is statistically insignificant at conventional levels for

[17] We find that if the proportion of peers in the top quartile increases relative to the middle quartile there is a significant decrease in student *i*'s endline reading test scores for male students (see Column 2, Panel A of Table 5).

the top peer quartile relative to the middle quartile.[18] For students in the top quartile we find that if you increase the proportion of peers in the top (bottom) quartile by 1 percentage point relative to the middle quartile, then endline reading tests scores decreases (increases) by 0.05 (0.06) points (see Column 8, Panel A of Table 3), although the effects are imprecisely estimated.[19] Finally, for students in the middle quartile the effect of changing the proportion of peers in the top (bottom) quartile relative to the middle quartile is much smaller in magnitude and statistically insignificant. The patterns for endline math test scores essentially mirror those found for endline reading test scores with the following exception. Unlike endline reading tests scores, for students in the bottom (middle) quartile we find that if the proportion of peers in the top quartile increases relative to the middle quartile there is an increase, not a decrease, in student *i*'s endline math test scores (see Columns 4 and 6, Panel B of Table 3), although the effects are imprecisely estimated.[20]

3.2.3 Testing the Models of Peer Effects

Thus far, we have focused on estimating the peer effects on student achievement. It is equally important to formally identify the underlying models of peer effects. As such, we test the predictions of the four models of peer effects outlined in Section 3.1: the monotonicity model (i.e., the effects of peers on student achievement is increasing in peer quality); the invidious

[18] For black students (see Column 4, Panel A of Table 7) however we find that a 1 percentage point increase in the proportion of peers in the top quartile relative to the middle is associated with a 0.15 point increase in student i's endline reading test scores (the effect is insignificant at conventional levels).

[19] For female students (see Column 8, Panel A of Table 4) however we find that if the proportion of peers in the bottom of the distribution increases relative to the middle there is a decrease in student *i*'s endline reading test scores, although the effects are imprecisely estimated.

[20] For black students (see Column 8, Panel A of Table 7) and Hispanic students (see Column 8, Panel A of Table 8) in the top quartile we find that a 1 percentage point increase in the proportion of peers in the top quartile relative to the middle is associated with a 0.16 and 0.05 point increase, respectively, in student *i*'s endline math test scores and for Hispanic students (see Column 8, Panel A of Table 8) in the top quartile we find that a 1 percentage point increase in the proportion of peers in the bottom quartile relative to the middle is associated with a 0.08 point decrease in student *i*'s endline math test scores. The effects however are insignificant at conventional levels.

comparison model (i.e., higher ability peers adversely influence the outcomes of students who are moved to a lower position in the achievement distribution), ability grouping (boutique) model (i.e., student performance is highest when their peers are similar to themselves), and frame of reference model (i.e., higher ability peers adversely influence the outcomes of students due to a lower academic self-concept).

Specifically, based on the nonlinear results for the full sample where both own student and peer effects are allowed to vary by placement in the subject specific achievement distribution (i.e., Columns 4, 6, and 8 of Table 3), we count the number of tests that are significant at the 10% level in the direction predicted by the model under question, the number of tests that are significant at the 10% level in the opposite direction predicted by the model under question, and the number of tests that are statistically insignificant. This is very similar to the inference procedure in Imberman, Kugler, and Sacerdote (2012).[21] We take any tests indicating significant in the correct (opposite) direction to be consistent (inconsistent) with the model. If there are no significant tests, irrespective of the direction predicted by the model, we find no support for the model. The results are presented in Table 9.

We find evidence against both the weak and strong monotonicity models for reading test scores. For math test scores while we also find evidence against the strong monotonicity model we do not find evidence for or against the weak monotonicity model (i.e., no tests are significant in the direction of or in the opposite direction the model suggest). For instance, consider the tests for the strong monotonicity model, 1 (2) out of 6 tests are significant in the opposite direction the model suggests for reading (math) test scores, while there are no significant tests in the direction predicted by the model. The evidence with respect to the ability grouping model is mixed.

[21] We similarly test the peer effect models for each sub-group (see Appendix Table A4). The patterns for the subgroups mostly coincide with those of the full sample.

Specifically, we find evidence against the ability grouping model based on math test scores (i.e., 1 out of 6 tests are significant in the opposite direction the model suggests) and evidence for the ability grouping model based on reading test scores (i.e., 2 out of 6 tests are significant in the direction the model suggests). We do however find stronger support for both the invidious comparison model and the frame of reference model. Specifically, we never find significant tests in the opposite direction predicted by either invidious comparison model or the frame of reference models. For the invidious comparison model, we find 0 (2) out of 4 tests are significant in the direction the model suggests for reading (math) test scores. For the weak (strong) frame of reference model, 1 (1) out of 3 (6) tests predict the model in the correct direction for reading test scores and 0 (2) out of 3 (6) tests predict the model in the correct direction for math test scores.

We also offer some suggestive evidence to further support the patterns above. Specifically, if we break up the insignificant results into those that predict the model in the correct and the incorrect direction, we find that all the tests go in the opposite direction predicted by the weak monotonicity models for both reading and math test scores. While the same is true for the strong monotonicity model for reading, for math all but 2 of the 6 tests go in the opposite direction predicted by the model. For the invidious comparison model we find that 4 (2) out of the 4 tests go in the direction predicted by the model for reading (math) test scores.   For the ability grouping model we find that 3 (4) out of 6 tests go in the opposite direction predicted by the model for reading (math) test scores. It is also important to note that the evidence for the ability grouping model only comes from the bottom quartiles of the reading achievement distribution both in terms of sign and statistical significance. Finally, for the weak frame of reference model all the tests go in the direction predicted by the model irrespective of test subject while for the strong frame of reference model all (4) of the 6 tests go in the direction predicted by the model for reading (math) test scores.[22]

Taken together, we appear to find no support for the weak and strong monotonicity models and little evidence in favor of the ability grouping model. However, we appear to find stronger support for the invidious comparison and frame of reference models. Our study is not the first one in the economics literature to present evidence supporting the frame of reference model and/or invidious comparison model. For instance, Pop-Eleches and Urquiola (2013), using survey data from Romanian secondary schools, show that children admitted into more selective schools by scoring just above a cut-off point perform worse potentially due to a reduction in their confidence and self-esteem due to their exposure to better peers. Bui, Craig, and Imberman (2013) also find some evidence supporting the invidious comparison and/or the frame of reference models in magnet schools. Specifically, the authors provide tentative evidence that students who are marginally eligible to enroll in magnet schools and therefore are exposed to higher achieving peers then they have would been in a regular public school tend to perform worse in terms of their own achievement. Finally, Lavy, Silva, and Weinhardt (2013), using survey data from England, propose the frame of reference and/or invidious comparison model as a potential explanation for the negative impact of high performing male peers on the achievement of male students.

Finally our findings may also shed some light on the channels through which peer effects in primary school operate. It may be the case that high quality peers in the classroom depresses

[22] While we get similar results based on reading test scores if we allow for interdependence between tests (i.e., do the Bonferroni correction), for math test scores the results with the Bonferroni correction do not allow us to either reject or support any of the peer effect models (see Appendix Table A5). However, we argue that this is largely an artifact of small sample sizes combined with the marginal level of significance found for the math test scores. Moreover, it is a well known fact that the Bonferroni tests are too conservative and may lack power for correlated tests.

the academic performance of all students presumably through the frame of reference model or the invidious comparison model. In this case, negative peer effects result from the interactions across students. Alternatively, as noted in Lavy, Paserman, and Schlosser (2012), an increase in the overall achievement of the classroom may force teachers to raise the level more towards higher ability students and this may hurt some students. The negative peer effects here result from a change in teaching practices and methods. With our data, it is not possible to directly see whether the teacher raises the level of teaching more towards higher ability students or not. That said, however, if the peer effects were to stem from changes in teachers' pedagogical practices, one would not observe any negative peer effects at the top of the achievement distribution and no effect whatsoever among the middle achievement students as we do. In other words, a teacher raising the bar so high that even students in the top quartile are hurt does not appear to be a valid explanation because for this to be the case we would also observe negative peer effects for middle achievement students and we do not.

3.2.4 Can We Reconcile Our Results with the Previous Literature?

Even though there is some evidence in the literature supporting our results pertaining to the peer effect models, our findings presented so far appear to be generally at odds with those found in the existing peer effect studies. What can account for this divergence in results? One possibility is it may be because of the differences in our estimation sample. We are focusing on students from disadvantaged neighborhoods and, as noted at the outset of the paper, peer interactions may differ by socio-economic status and family background (see for example, Ludwig, Duncan and Hirsfield 2001). Alternatively, it may be because of the differences in the level of peer aggregation (i.e., classroom-level vs. grade-level). Finally, it may be because of the differences in the peer effect identification strategies. Even though the existing peer effect literature based on grades K through 12 carefully addresses the identification problems in the estimation of peer effects, they may not be able to fully account for all the potential confounding effects in survey data.

Ideally, we would like to test the sensitivity of our results based on all three potential explanations. Due to the nature of our dataset, however, we cannot say much about what the peer effect estimates would be if we had used the same identification strategy and the same peer measure in a nationally representative sample of students. That being said, we can still exploit the unique features of the TTTDR data set to shed some light on how our use of classroom level peer achievement and our use of randomized data affects our findings.

As discussed Section 2.1, 10 schools were present in the study for both 2004-2005 and 2005-2006 with a total of 760 student observations (see Appendix Table A6 for summary statistics). Within these schools, teachers were either retained from the first year or new teachers were added in the second year. For the purpose of this analysis it is also important to recall that teachers who were retained in the second year continue to teach the same grade they taught in the first year but have new classrooms of randomly assigned students. The repeated cross-section nature of the subset of TTTDR data set allows us to identify the peer effects from cohort-to-cohort variation in peer achievement within the same grade and school. Specifically, we estimate the following equation

$$TS_{icgst}^{end} = \beta_0 + \beta_1 \overline{TS}_{-i,cs}^{base} + \beta_2 \overline{TS}_{icgst}^{base} + SC_{icgst}^{'}\delta + TC_{cgst}^{'}\gamma + \alpha_g + \theta_s + \lambda_{gt} + \nu_{icgst} \tag{6}$$

where *i* denotes individuals, *c* denotes the classroom, *g* denotes the grade (block), *s* denotes school, *t* denotes time α_g , θ_s , λ_{gt} are grade, school, and grade by year fixed effects respectively; β_1 represents the effect of average peer achievement on student achievement, and all other variables are as previously defined. For comparative reasons, we re-estimate equation (3) for the same subset of the TTTDR data set.

In this simple set-up, any potential divergence in the coefficient estimates of average peer achievement from equation (3) and equation (6) are likely to be a by-product of non-random sorting of students (i.e., pre-existing trends). Columns 1 and 2 of Panel A (B) of Table 10 present the reading (math) classroom level peer achievement estimates from equations (3) and (6), respectively. The peer effect coefficient estimates for reading and math test scores from the specification where we use the cohort-to-cohort variation (randomization) in peer achievement are -0.398 (-0.437) and -0.221 (-0.277), respectively. While using the cohort-to-cohort identification strategy reduces the magnitudes of both the reading and math coefficient estimates relative to using the randomized nature of the data, the discrepancy between them is not large enough to rule out random statistical error. As such, it appears that there is no compelling evidence suggesting that identification of peer effects by using cohort-to-cohort variation in peer achievement within the same grade and school generates biased coefficients.

Next we replace classroom level peer achievement with the grade (block) level peer achievement in equation (6). In this revised set up, any potential discrepancy between the peer effect estimates using peer achievement at the classroom level and the grade level would reflect differences in level of peer aggregation.[23] Column (3) of Panel A (B) in Table 10 displays the peer effect estimates for reading (math) test scores using grade level peer achievement in equation (6). The coefficient estimate on reading peer achievement is about half the size of the specification where we use classroom level peer achievement (Column 2 of Panel A in Table 10). As for math, the corresponding coefficient (Column 3 of Panel B in Table 10) is again

[23] It is important to note that we are able to compare the estimates based on classroom level peer achievement and grade (block) level peer achievement given students are randomly assigned at the grade level across classrooms.

smaller in magnitude but the decrease in the coefficient estimate is smaller than the one we observe for reading achievement.[24] These results suggest that the importance of identifying the salient peer group in estimating peer effects cannot be understated.

Stepping back and viewing the complete exercise from this section our focus on students in primary schools in *disadvantaged* neighborhoods may serve as an explanation for the discrepancy between our results and those found in the previous literature given neither the potentially confounding effects in repeated cross-section data nor the use of a classroom level peer measure (as opposed to a grade level measure) appear to be valid explanations.

3.2.5 Robustness Checks

We undertake several sensitivity checks to examine the robustness of our results. First, following Foster (2006) we replace the average baseline peer achievement with the median baseline achievement level of the classroom and re-run all the specifications. The results from this exercise are qualitatively similar to those presented in the paper. Second, we added the standard deviation of the baseline peer achievement along with average baseline peer achievement. The coefficient estimates on the dispersion measure are not different than zero in any of the specifications. Third, we tried including average peer reading and math achievement scores simultaneously to subject-specific achievement equations. Average peer effect coefficients for both subjects on endline scores remain almost intact. They are, however, less precisely estimated. This may not be surprising given the high correlation between these two peer achievement measures.

[24] Ideally, we would like to extend these two exercises to own achievement groups but a further break down of the repeated cross-section sample leads to a very limited number of observations and a limited number of blocks. As such, random statistical error is likely to contribute to the variation in the peer effect estimates. Nevertheless, the patterns are qualitatively similar if re-estimate equation (6) by own achievement to those presented in the paper.

Fourth, to examine the potential differential effects of peer quality at different grades, we divide the sample into lower grades (kindergarten and first grade) and upper grades (second through fifth). The peer effect coefficient estimates from the lower versus the upper grades do not indicate any discernible pattern. Fifth, rather than splitting the sample within each achievement group based on selected student characteristics, we run fully interacted models (e.g., the female indicator variable interacted with all covariates) within each achievement group. We also repeated a similar exercise within each subgroup and run fully interacted models in ability (i.e., baseline achievement indicator variables—top, middle, and bottom—interacted with all covariates). The precision of our results from these robustness checks are very similar to those presented in the paper. Sixth, we choose different cut-off points to describe the bottom and top achievement groups (i.e., one-third). Doing so does not alter our conclusions. Finally, the results are similar, although slightly more precisely estimated, if we instead cluster at the classroom-level. All these results are available upon request.

4. Conclusion

For decades, there has been a flurry of research by social scientists trying to pinpoint the underlying determinants of student achievement, particularly since the Coleman Report was released in 1966. Despite this, we still know very little about which specific education policies are more effective at improving student achievement outcomes. This paper further analyzes how to improve student achievement with a particular interest on the effect of peers on student achievement.[25]

We use data from a well-executed randomized experiment which allows us to measure the causal effect of peer quality, as well as affords us a large sample of primary schools, students,

[25] In a companion paper, we examine in details the contextual peer effects (see Antecol, Eren, and Ozbeklik 2013).

teachers, and states. Furthermore, our data comes from a disadvantaged part of the student population which allows us to take a closer look at peer effects in a setting where the influence of family background may be particularly less pronounced and peer dynamics might differ from those in a nationally representative sample of students given students in this population may be particularly sensitive to peer interactions.

Unlike the existing literature which generally finds positive and significant effects or small positive to no effects, we find that the average classroom baseline peer achievement *adversely* influences student's own endline achievement. The linear-in-means model, however, masks a great deal of information. We therefore extend our analysis to take into account nonlinearities in peer effects which reveals substantial heterogeneity across the achievement distribution. Specifically, we consistently find negative peer effects at the bottom of the reading achievement distribution for the full sample, as well as for all subgroups, although some effects are imprecisely estimated. We also find negative peer effects at the top of the reading achievement distribution for the full sample and all subgroups however for certain subgroups the effects are more precisely estimated (particularly at the top of the distribution). Peer effects estimates on reading achievement for middle ability students, on the other hand, are essentially zero for the full sample and across all subgroups of interest. Turning to math test scores, we find that peer quality adversely affects student achievement for the full sample over the entire achievement distribution, although the effects are imprecisely estimated. The full sample pattern for math achievement is essentially mirrored for two subgroups only. Specifically, we observe negative (and significant) peer effects for male students at almost all achievement levels and for Hispanic students in the middle achievement group. Taken altogether, direct peer effects as opposed to teacher responses to student compositional changes appear to be driving our results.

We also test several peer effect models to further our understanding of the peer interactions in primary schools in disadvantaged neighborhoods. While we find no evidence to support the monotonicity model and little evidence in favor of the ability grouping model, we find stronger evidence to support the frame of reference and the invidious comparison models. Furthermore, in an attempt to reconcile our results with the existing literature we use a unique feature of our data to investigate how sensitive our results are to our use of random data for identification of peer effects and to differences in the level of peer aggregation. We find suggestive evidence that our focus on students in primary schools in *disadvantaged* neighborhoods may explain the discrepancy between our results and those found in the existing literature as neither the potentially confounding effects in repeated cross-section data nor the use of a classroom level peer measure (as opposed to a grade level measure) change our peer effect coefficients enough to be valid explanations. More research focusing on peer interactions in disadvantaged schools would be beneficial in understanding the peer effect dynamics in a setting where the problems with the education system in the United States are most evident and arguably more important from a policy perspective.

Finally, we performed a simple policy experiment similar to that of Lavy, Silva, and Weinhardt (2012) where we take an average classroom in our sample and rank students in the classroom from the best to the worst based on their baseline achievement. We then group all of the students with below-median reading (math) test scores in a "low achievement" classroom and the remaining students in a "high achievement" classroom. The manipulation of the average classroom in this way has the following effects: (1) the proportion of students in the bottom quartile of the baseline reading (math) achievement doubles from 24 (25) percent to 48 (50) percent in the low achievement reading (math) classroom and decreases from 24 (25) percent to

0 in the high achievement reading (math) classroom; and (2) the proportion of students in the top quartile of the baseline reading (math) achievement decreases from 26 (25) percent to 0 in the low achievement reading (math) classroom and doubles from 26 (25) percent to 52 (50) percent in the high achievement reading (math) classroom. Using our results in Table 3 we find that this tracking policy appears to benefit students in the low achievement classroom and to hurt the students in the high achievement classroom. It is important to note however that the predictions of a large policy intervention like the one in our policy experiment should be viewed with caution given the potential unintended consequences (i.e., endogeneous sorting) of the intervention (see Carell, Sacerdote, and West 2013).

References

Antecol, Heather, Ozkan Eren, and Serkan Ozbeklik. 2013. “Gender and Racial Composition of Peers and Student Achievement in Disadvantaged Primary Schools*,” Unpublished Manuscript*.

Arcidiacono, Peter, and Sean Nicholson. 2005. “Peer Effects in Medical School,” *Journal of Public Economics*, 89(2-3): 327-350.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. “Social Preferences and the Response to Incentives: Evidence from Personnel Data,” *Quarterly Journal of Economics*, 120(3): 917–62.

Bertrand, Marianne. 2010. “New Perspectives in Gender,” in *Handbook of Labor Economics* Volume 4B, eds. Orley Ashenfelter and David Card, 1545-1592, Elsevier.

Betts, Julian R., and Andrew Zau. 2004. “Peer Groups and Academic Achievement: Panel Evidence from Administrative Data,” Unpublished Manuscript.

Boozer, Michael A., and Stephen E. Cacciola. 2001. “Inside the Black Box of Project Star: Estimation of Peer Effects,” Economic Growth Center Discussion Paper No. 832.

Brown, Jennifer. 2011. “Quitters Never Win: The (Adverse) Incentive Effects of Competing with Superstars,” *Journal of Political Economy*, 119(5): 982-1013.

Bui, Sa A., Steven G. Craig and Scott A. Imberman. (2013). “Is Gifted Education a Bright Idea? Assessing the Impact of Gifted and Talented Programs on Students,” *American Economic Journal: Economic Policy*, forthcoming.

Burke, Mary A., and Tim R. Sass. 2013. “Classroom Peer Effects and Student Achievement,” *Journal of Labor Economics*, 31(1): 51-82.

Carrell, Scott E., Richard L. Fullerton, and James E. West. 2009. “Does Your Cohort Matter? Measuring Peer Effects in College Achievement,” *Journal of Labor Economics*, 27(3): 439-464.

Carrell, Scott E. and Mark L. Hoekstra. 2010. “Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids,” *American Economic Journal: Applied Economics*, 2(1):211-228.

Carell, Scott E., Bruce I. Sacerdote, and James E. West. 2013. “From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation,” *Econometrica*, 81(3): 855-882.

Carman, Katherine G., and Lei Zhang. 2012. “Classroom Peer Effects and Academic Achievement: Evidence from a Chinese Middle School,” *Chinese Economic Review*, 23, 223-237.

Case, Anne, and Larry F. Katz. 1991. "The Company You Keep: The Effects of Family and Neighborhood on Disadvantaged Youths," NBER Working Paper No. 3705.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources*, 41(4): 778-820.

Coleman, James.S. 1966. "Equality of Educational Opportunity," Washington: U.S. Government Printing Office, 1966 [summary report].

Constantine, Jill, Daniel Player, Tim Silva, Kristin Hallgren, Mary Grider, John Deke, and Elizabeth Warner. 2009. "An Evaluation of Teachers Trained Through Different Routes to Certification: Final Report," NCEE 2009-4043. Institute of Education Sciences. Department of Education.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences," *Journal of Economic Literature*, 47(2): 1-27.

Curto, Vilsa E., and Roland G. Fryer, Jr. 2013. "The Potential of Urban Boarding Schools for the Poor: Evidence from SEED," *Journal of Labor Economics*, forthcoming.

Duflo, Esther, Pascaline Dupas and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101(5): 1739-1774.

Epple, Dennis, and Richard E. Romano. 2011. "Peer Effects in Education: A Survey of the Theory and Evidence," in *Handbook of Social Economics* Vol. 1, eds. Jess Benhabib, Alberto Bisin and Matthew O. Jackson, 1053-1163. Amsterdam, North-Holland.

Falk, Armin, and Andrea Ichino. 2006. "Clear Evidence on Peer Effects," *Journal of Labor Economics* 24(1): 39–57.

Fleischman, Howard L., Paul J. Hopstock, Pelczar, Marisa P. Pelczar, and Brooke E. Shelley. 2010. "Highlights from PISA 2009: Performance of U.S. 15-Year-Old Students in Reading," Mathematics, and Science Literacy in an International Context. NCES 2011-004.

Frölich, Markus, and Blaise Melly. 2013. "Unconditional Quartile Treatment Effects under Endogeneity," *Journal of Business and Economic Statistics*, 31(3): 346-357.

Foster, Gigi. 2006. "It's Not Your Peers, and It's Not Your friends: Some Progress toward Understanding the Educational Peer Effect Mechanism," *Journal of Public Economics*, 90(10-11): 1455-1475.

Gaviria, Alejandro and Steven Raphael. 2001. "School-Based Peer Effects and Juvenile Behavior," *Review of Economics and Statistics*, 83(2): 257–268.

Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo. 2009. “Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments,” *American Economic Journal: Applied Economics*, 1(4): 34-68.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. “Does Peer Ability Affect Student Achievement,” *Journal of Applied Econometrics*, 18(5): 527-544.

Hanushek, Eric A. 2006. “School Resources,” in *Handbook of the Economics of Education* Vol. 2, eds. Eric A Hanushek and Finis Welch, 865-908, Elsevier.

Hanushek, Eric A., and Margaret E. Raymond. 2005. “Does School Accountability Lead to Improved Student Performance?” *Journal of Policy Analysis and Management*, 24(2): 297-327.

Hanushek, Eric A., and Steven G. Rivkin. 2006. “Teacher Quality,” in *Handbook of the Economics of Education* Vol. 2, eds. Eric A. Hanushek and Finis Welch, 1051-1162, Elsevier.

Hanushek, Eric A. and , Steve G. Rivkin. 2009. “Harming the Best: How Schools Affect the Black-White Achievement Gap,” *Journal of Policy Analysis and. Management*, 28 (Summer): 366–393.

Hoxby, Caroline M. 2000. “Peer Effects in the Classroom: Learning from Gender and Race Variation,” NBER Working Paper No. 7867.

Hoxby, Caroline M. and Gretchen Weingarth. 2006. “Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects,” unpublished manuscript.

Imberman, Scott A., Adriana D. Kugler, and Bruce I. Sacerdote. 2012. “Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees,” *American Economic Review*, 102(5): 2048-2082.

Jackson, Kirabo C. 2013. “Can Higher Achieving Peers Explain the Benefits to Attending Selective Schools? Evidence from Trinidad Tobago,” *Journal of Public Economics* 108(6): 63-77.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. “Identifying Effective Classroom Practices Using Student Achievement Data,” *Journal of Human Resources*, 46(3): 587-613.

Kling, Jeffrey R., Jens Ludwig, and Larry F. Katz. 2005. “Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment,” *Quarterly Journal of Economics*, 120(1): 87–130.

Krueger, Alan B., and Diane M. Whitmore. 2001. “The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Results: Evidence from Project STAR,” *Economic Journal*, 111(468): 1-28.

Lavy, Victor, Daniele Paserman, and Analia Schlosser. 2012. "Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom," *Economic Journal*, 122(559): 208-237.

Lavy, Victor, Olma Silva, and Felix Weinhardt. 2012. "The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools," *Journal of Labor Economics*, 30(2): 367-414.

Ludwig, Jens, Greg J. Duncan, and Paul Hirschfield. 2001. "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility," *Quarterly Journal of Economics,* 116(2): 655-679.

Lyle, David S. 2007. "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point," *Review of Economics and Statistics,* 89(2): 289-299.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies,* 60(3): 531-542.

Marsh, Herbert W., and Parker, John W. 1984. "Determinants of Student Self-Concept: Is it Better to be a Relatively Large Fish in a Small Pond Even If You Don't Learn to Swim as Well?" *Journal of Personality and Social Psychology*, 47(1): 213-231.

Marsh, Herbert W., 1987. "The Big Fish Little Pond Effect on Academic Self-Concept," *Journal of Educational Psychology*, 79(3): 280-295.

Marsh, Herbert W., and Hau, Kit-Tai. 2003. "Big Fish Little Pond Effect on Academic Self-Concept: A Cross Cultural (26-Country) Test of the Negative Effects of Academic Selective Schools," *American Psychologist*, 58(5): 364-376.

Mass, Alexandre and Enrico Moretti. 2009. "Peers at Work," *American Economic Review*, 99(1): 112-145.

Moffitt, Robert A. 2001. "Policy Interventions, Low-Level Equilibria, and Social Interactions," in *Social Dynamics*, eds. Steven N. Durlauf and H. Peyton Young, 45-82. Cambridge, MA: MIT Press.

Pop-Eleches, Christian and Miguel Urquiola. 2013. "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, 103(4): 1289-1324.

Sass, Tim R., Jane Hannaway, Zeyu Xu, David N. Figlio and Li Feng. 2012. "Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools," *Journal of Urban Economics*, 72(2-3): 104-122.

Sacerdote, Bruce I. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates," *Quarterly Journal of Economics*, 116(2): 681-704.

Sacerdote, Bruce I. 2011. "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" in *Handbook of the Economics of Education* Vol. 3, eds. Erik Hanushek, Stephen Machin and Ludger Woessmann, 249-277. Elsevier.

Stinebrickner, Ralph, and Todd R. Stinebrickner. 2006. "What Can Be Learned about Peer Effects Using College Roommates? Evidence from New Survey Data and Students from Disadvantaged Backgrounds," *Journal of Public Economics*, 90(8-9): 1435-1454.

US Dept of Education, 2009. Race to the Top Executive Summary. Retrieved from http://www2.ed.gov/programs/racetothetop/

Vigdor, Jacob L. 2006. "Peer Effects in Neighborhoods and Housing," in *Deviant Peer Influences in Programs for Youth: Problems and Solutions*, eds. Kenneth A. Dodge., Thomas J. Dishion, Jennifer E. Lansford, Guilford Press.

Vigdor, Jacob L., and Thomas Nechyba. 2007. "Peer Effects in North Carolina Public Schools," in *Schools and the Equal Opportunity Problem*, eds. Ludger Woessmann and Paul E. Peterson, 73-102. Cambridge, MA: MIT Press.

Zimmerman, David J. 2003. "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment," *Review of Economics and Statistics*, 85(1): 9-23.

Table 1: Student Summary Statistics and Basic Randomization Regressions

	TTTDR Students	AC Students	TC (Control) Students	Coefficient (Standard Error)
	Mean (Standard Error)	Mean (Standard Error)	Mean (Standard Error)	**AC**
Endline Reading Test Score (NCE)	38.59	39.07	38.13	…..
	(20.21)	(20.38)	(20.03)	
Endline Math Test Score (NCE)	42.43	42.59	42.28	…..
	(22.66)	(22.83)	(22.51)	
Baseline Reading Test Score (NCE)	38.93	39.86	38.05	0.00
	(20.87)	(20.92)	(20.80)	(0.00)
Baseline Math Test Score (NCE)	42.55	42.93	42.18	-0.00
	(21.28)	(21.06)	(21.49)	(0.00)
Female (1=Yes)	0.45	0.46	0.44	0.01
	(0.49)	(0.49)	(0.49)	(0.01)
Race				
White	0.09	0.09	0.08	-0.02
	(0.28)	(0.29)	(0.28)	(0.03)
Black	0.35	0.34	0.35	-0.03
	(0.47)	(0.47)	(0.47)	(0.02)
Hispanic	0.47	0.47	0.47	0.01
	(0.49)	(0.49)	(0.49)	(0.02)
Free/Reduced Lunch (%)	0.75	0.74	0.77	-0.04
	(0.42)	(0.43)	(0.41)	(0.03)
Sample Size	2,610	1,280	1,340	

NOTES: All test scores are expressed in NCEs. NCE scale has a mean 50 and standard deviation 21.06 nationally. Randomization regression tests control for block fixed effects. The standard errors clustered at the block level are reported. AC indicator takes the value of one if the student is taught by a AC teacher and takes the value of zero if the student is taught by a TC teacher. Due to confidential nature of the data, the sample sizes are rounded to the nearest tenth.

Table 2: Randomization Tests

Dependent Variable: Own Baseline Test Scores	Coefficients (Standard Error)			
	Reading	Math	Reading	Math
Average Peer Baseline Reading Achievement	-0.864*** (0.169)	…..	0.018 (0.034)	…..
Average Peer Baseline Math Achievement	…..	-0.722*** (0.168)	…..	-0.063 (0.053)
Average Block Baseline Reading Achievement	…..	…..	-25.478*** (0.716)	…..
Average Block Baseline Math Achievement	…..	…..	…..	-25.783*** (0.754)

NOTES: All test scores are expressed in NCEs. Standard errors clustered at the block level are reported. Randomization regressions control for block fixed effects. Average peer subject-specific baseline achievement measured at the classroom level.
** significant at 5%, *** significant at 1%.

Table 3: Estimates of Peer Effects by Own Student Baseline Achievement at the Grade Level

Dependent Variable: Endline Test Scores	Coefficients (Standard Error)							
Panel A: Reading Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.182***		-0.453***		-0.067		-0.269	
	(0.069)		(0.164)		(0.089)		(0.196)	
Proportion of Peers in Top 25%		-0.048		-0.060		-0.029		-0.050
		(0.032)		(0.086)		(0.040)		(0.077)
Proportion of Peers in Bottom 25%		0.047		0.140***		0.022		0.061
		(0.030)		(0.047)		(0.046)		(0.080)
Own Baseline Test Score	0.632***	0.634***	0.544***	0.562***	0.726***	0.728***	0.552***	0.564***
	(0.020)	(0.020)	(0.080)	(0.080)	(0.057)	(0.057)	(0.052)	(0.052)
Bottom vs. Middle (p-value)	0.02							
Bottom vs. Top (p-value)	0.44							
Middle vs. Top (p-value)	0.33							
Top Proportion vs. Bottom Proportion (p-value)		0.03		0.04		0.40		0.31
Sample Size	2,610		640		1,290		680	
Panel B: Math Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75%**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.236**		-0.281		-0.194		-0.198	
	(0.103)		(0.181)		(0.119)		(0.192)	
Proportion of Peers in Top 25%		0.018		0.058		0.052		-0.032
		(0.036)		(0.086)		(0.054)		(0.089)
Proportion of Peers in Bottom 25%		0.115**		0.117		0.089*		0.160*
		(0.045)		(0.083)		(0.051)		(0.093)
Own Baseline Test Score	0.627***	0.630***	0.555***	0.565***	0.654***	0.661***	0.520***	0.530***
	(0.020)	(0.019)	(0.068)	(0.066)	(0.072)	(0.072)	(0.080)	(0.080)
Bottom vs. Middle (p-value)	0.67							
Bottom vs. Top (p-value)	0.73							
Middle vs. Top (p-value)	0.99							
Top Proportion vs. Bottom Proportion (p-value)		0.09		0.62		0.61		0.13
Sample Size	2,580		670		1,250		660	

NOTES: All test scores are expressed in NCEs. Standard errors are clustered at the block level. All specifications control for block fixed effects. Student controls include gender, race/ethnicity and eligibility for free lunch. Teacher controls include teacher's type: AC or TC, gender, race/ethnicity and teaching experience. Average peer subject-specific achievement measured at the classroom level. The proportion of top 25% and bottom 25% of peers in a classroom are based on the grade and subject-specific baseline test score distribution. Due to confidential nature of the data, the sample sizes are rounded to the nearest tenth.

* significant at 10%, ** significant at 5%, *** significant at 1%.

Table 4: Estimates of Peer Effects by Own Student Baseline Achievement at the Grade Level for Female Students

Dependent Variable: Endline Test Scores	Coefficients (Standard Error)							
Panel A: Reading Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.118		-0.428		-0.025		-0.207	
	(0.094)		(0.265)		(0.153)		(0.247)	
Proportion of Peers in Top 25%		-0.012		-0.071		0.006		-0.085
		(0.037)		(0.124)		(0.055)		(0.099)
Proportion of Peers in Bottom 25%		0.048		0.133*		0.037		-0.036
		(0.041)		(0.079)		(0.086)		(0.109)
Own Baseline Test Score	0.608***	0.609***	0.562***	0.573***	0.623***	0.621***	0.452***	0.460***
	(0.023)	(0.023)	(0.112)	(0.113)	(0.080)	(0.081)	(0.076)	(0.075)
Bottom vs. Middle (p-value)	0.18							
Bottom vs. Top (p-value)	0.53							
Middle vs. Top (p-value)	0.52							
Top Proportion vs. Bottom Proportion (p-value)		0.27		0.16		0.76		0.73
Sample Size	1,190		290		580		330	

Panel B: Math Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75%**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.037		0.144		0.009		-0.343	
	(0.124)		(0.221)		(0.182)		(0.332)	
Proportion of Peers in Top 25%		0.030		0.129		0.067		-0.158
		(0.057)		(0.161)		(0.089)		(0.158)
Proportion of Peers in Bottom 25%		0.074		0.009		0.051		0.146
		(0.062)		(0.138)		(0.087)		(0.168)
Own Baseline Test Score	0.641***	0.638***	0.625***	0.613***	0.641***	0.636***	0.552***	0.563***
	(0.027)	(0.027)	(0.099)	(0.009)	(0.124)	(0.123)	(0.128)	(0.127)
Bottom vs. Middle (p-value)	0.62							
Bottom vs. Top (p-value)	0.22							
Middle vs. Top (p-value)	0.36							
Top Proportion vs. Bottom Proportion (p-value)		0.60		0.57		0.89		0.18
Sample Size	1,170		310		580		280	

NOTES: All test scores are expressed in NCEs. Standard errors are clustered at the block level. All specifications control for block fixed effects. Student controls include race/ethnicity and eligibility for free lunch. Teacher controls include teacher's type: AC or TC, gender, race/ethnicity and teaching experience. Average peer subject-specific achievement measured at the classroom level. The proportion of top 25% and bottom 25% of peers in a classroom are based on the grade and subject-specific baseline test score distribution. Due to confidential nature of the data, the sample sizes are rounded to the nearest tenth.

* significant at 10%, ** significant at 5%, *** significant at 1%.

Table 5: Estimates of Peer Effects by Own Student Baseline Achievement at the Grade Level for Male Students

Dependent Variable: Endline Test Scores	Coefficients (Standard Error)							
Panel A: Reading Test Scores				**Own Student Baseline Achievement**				
	All		**Bottom 25%**		**Middle 25%-75**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.224**		-0.446**		-0.171		-0.596*	
	(0.094)		(0.198)		(0.133)		(0.344)	
Proportion of Peers in Top 25%		-0.080**		-0.092		-0.063		-0.057
		(0.040)		(0.121)		(0.061)		(0.118)
Proportion of Peers in Bottom 25%		0.055		0.135*		0.043		0.142
		(0.040)		(0.079)		(0.060)		(0.133)
Own Baseline Test Score	0.642***	0.643***	0.498***	0.515***	0.759***	0.765***	0.535***	0.572***
	(0.028)	(0.027)	(0.127)	(0.123)	(0.082)	(0.079)	(0.077)	(0.073)
Bottom vs. Middle (p-value)	0.25							
Bottom vs. Top (p-value)	0.70							
Middle vs. Top (p-value)	0.24							
Top Proportion vs. Bottom Proportion (p-value)		0.01		0.11		0.21		0.26
Sample Size	1,420		360		710		350	
Panel B: Math Test Scores				**Own Student Baseline Achievement**				
	All		**Bottom 25%**		**Middle 25%-75%**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.441***		-0.574**		-0.360**		-0.454	
	(0.130)		(0.241)		(0.180)		(0.299)	
Proportion of Peers in Top 25%		-0.011		0.010		0.062		-0.041
		(0.049)		(0.120)		(0.073)		(0.138)
Proportion of Peers in Bottom 25%		0.151***		0.186*		0.133*		0.260**
		(0.056)		(0.108)		(0.071)		(0.130)
Own Baseline Test Score	0.612***	0.620***	0.489***	0.512***	0.654***	0.666***	0.436***	0.460***
	(0.027)	(0.027)	(0.118)	(0.118)	(0.086)	(0.085)	(0.106)	(0.110)
Bottom vs. Middle (p-value)	0.48							
Bottom vs. Top (p-value)	0.74							
Middle vs. Top (p-value)	0.79							
Top Proportion vs. Bottom Proportion (p-value)		0.02		0.27		0.48		0.11
Sample Size	1,410		360		670		380	

NOTES: All test scores are expressed in NCEs. Standard errors are clustered at the block level. All specifications control for block fixed effects. Student controls include race/ethnicity and eligibility for free lunch. Teacher controls include teacher's type: AC or TC, gender, race/ethnicity and teaching experience. Average peer subject-specific achievement measured at the classroom level. The proportion of top 25% and bottom 25% of peers in a classroom are based on the grade and subject-specific baseline test score distribution. Due to confidential nature of the data, the sample sizes are rounded to the nearest tenth.
* significant at 10%, ** significant at 5%, *** significant at 1%.

Table 6: Estimates of Peer Effects by Own Student Baseline Achievement at the Grade Level for Free-Lunch Eligible Students

Dependent Variable: Endline Test Scores	Coefficients (Standard Error)							
Panel A: Reading Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.157*		-0.459**		0.005		-0.546**	
	(0.096)		(0.191)		(0.123)		(0.256)	
Proportion of Peers in Top 25%		-0.019		-0.045		0.003		-0.036
		(0.038)		(0.094)		(0.049)		(0.113)
Proportion of Peers in Bottom 25%		0.061*		0.146***		0.028		0.163
		(0.033)		(0.053)		(0.050)		(0.110)
Own Baseline Test Score	0.633***	0.634***	0.544***	0.563***	0.797***	0.797***	0.510***	0.544***
	(0.025)	(0.024)	(0.081)	(0.080)	(0.055)	(0.055)	(0.105)	(0.105)
Bottom vs. Middle (p-value)	0.05							
Bottom vs. Top (p-value)	0.77							
Middle vs. Top (p-value)	0.05							
Top Proportion vs. Bottom Proportion (p-value)		0.11		0.07		0.72		0.20
Sample Size	1,980		570		1,040		360	
Panel B: Math Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75%**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.218*		-0.243		-0.060		-0.236	
	(0.128)		(0.212)		(0.123)		(0.281)	
Proportion of Peers in Top 25%		0.016		0.038		0.079		-0.016
		(0.048)		(0.093)		(0.067)		(0.185)
Proportion of Peers in Bottom 25%		0.112**		0.086		0.078		0.185
		(0.049)		(0.091)		(0.051)		(0.122)
Own Baseline Test Score	0.639***	0.641***	0.538***	0.546***	0.619***	0.622**8	0.527***	0.539***
	(0.024)	(0.023)	(0.075)	(0.072)	(0.083)	(0.082)	(0.119)	(0.120)
Bottom vs. Middle (p-value)	0.45							
Bottom vs. Top (p-value)	0.97							
Middle vs. Top (p-value)	0.57							
Top Proportion vs. Bottom Proportion (p-value)		0.16		0.71		0.99		0.36
Sample Size	1,950		570		990		400	

NOTES: All test scores are expressed in NCEs. Standard errors are clustered at the block level. All specifications control for block fixed effects. Student controls include gender and race/ethnicity. Teacher controls include teacher's type: AC or TC, gender, race/ethnicity and teaching experience. Average peer subject-specific achievement measured at the classroom level. The proportion of top 25% and bottom 25% of peers in a classroom are based on the grade and subject-specific baseline test score distribution. Due to confidential nature of the data, the sample sizes are rounded to the nearest tenth.
* significant at 10%, ** significant at 5%, *** significant at 1%.

Table 7: Estimates of Peer Effects by Own Student Baseline Achievement at the Grade Level for Black Students

Dependent Variable: Endline Test Scores	Coefficients (Standard Error)							
Panel A: Reading Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.082		-0.319		0.051		-0.351	
	(0.108)		(0.313)		(0.141)		(0.291)	
Proportion of Peers in Top 25%		0.032		0.150		0.034		-0.010
		(0.056)		(0.137)		(0.069)		(0.118)
Proportion of Peers in Bottom 25%		0.072**		0.198**		0.047		0.110
		(0.032)		(0.082)		(0.075)		(0.153)
Own Baseline Test Score	0.652***	0.653***	0.747***	0.760***	0.678***	0.675***	0.460***	0.484***
	(0.028)	(0.028)	(0.124)	(0.122)	(0.100)	(0.101)	(0.102)	(0.099)
Bottom vs. Middle (p-value)	0.28							
Bottom vs. Top (p-value)	0.94							
Middle vs. Top (p-value)	0.21							
Top Proportion vs. Bottom Proportion (p-value)		0.55		0.76		0.89		0.53
Sample Size	900		180		500		220	

Panel B: Math Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75%**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.240		-0.367		-0.050		0.062	
	(0.186)		(0.310)		(0.208)		(0.489)	
Proportion of Peers in Top 25%		0.029		0.070		0.040		0.162
		(0.059)		(0.131)		(0.098)		(0.170)
Proportion of Peers in Bottom 25%		0.110		0.143		0.029		0.126
		(0.072)		(0.140)		(0.070)		(0.161)
Own Baseline Test Score	0.668***	0.672***	0.514***	0.520***	0.625***	0.629***	0.633***	0.614***
	(0.035)	(0.035)	(0.170)	(0.170)	(0.132)	(0.132)	(0.165)	(0.162)
Bottom vs. Middle (p-value)	0.39							
Bottom vs. Top (p-value)	0.45							
Middle vs. Top (p-value)	0.83							
Top Proportion vs. Bottom Proportion (p-value)		0.38		0.70		0.92		0.87
Sample Size	900		230		470		210	

NOTES: All test scores are expressed in NCEs. Standard errors are clustered at the block level. All specifications control for block fixed effects. Student controls include gender and eligibility for free lunch. Teacher controls include teacher's type: AC or TC, gender, race/ethnicity and teaching experience. Average peer subject-specific achievement measured at the classroom level. The proportion of top 25% and bottom 25% of peers in a classroom are based on the grade and subject-specific baseline test score distribution. Due to confidential nature of the data, the sample sizes are rounded to the nearest tenth.
* significant at 10%, ** significant at 5%, *** significant at 1%.

Table 8: Estimates of Peer Effects by Own Student Baseline Achievement at the Grade Level for Hispanic Students

Dependent Variable: Endline Test Scores	Coefficients (Standard Error)							
Panel A: Reading Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.080		-0.361		0.044		-0.204	
	(0.160)		(0.341)		(0.223)		(0.438)	
Proportion of Peers in Top 25%		-0.048		-0.096		-0.034		-0.054
		(0.051)		(0.122)		(0.069)		(0.121)
Proportion of Peers in Bottom 25%		0.048		0.135*		0.017		0.055
		(0.056)		(0.074)		(0.079)		(0.191)
Own Baseline Test Score	0.621***	0.618***	0.480***	0.497***	0.812***	0.806***	0.402***	0.416***
	(0.032)	(0.031)	(0.101)	(0.099)	(0.081)	(0.080)	(0.112)	(0.113)
Bottom vs. Middle (p-value)	0.32							
Bottom vs. Top (p-value)	0.77							
Middle vs. Top (p-value)	0.61							
Top Proportion vs. Bottom Proportion (p-value)		0.20		0.10		0.62		0.62
Sample Size	1,230		400		610		220	
Panel B: Math Test Scores			**Own Student Baseline Achievement**					
	All		**Bottom 25%**		**Middle 25%-75%**		**Top 25%**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Peer Baseline Achievement	-0.205		-0.219		-0.370**		0.365	
	(0.128)		(0.300)		(0.185)		(0.450)	
Proportion of Peers in Top 25%		-0.012		0.063		-0.041		0.049
		(0.057)		(0.156)		(0.095)		(0.256)
Proportion of Peers in Bottom 25%		0.089		0.094		0.152*		-0.079
		(0.065)		(0.128)		(0.085)		(0.245)
Own Baseline Test Score	0.609***	0.611***	0.551***	0.562***	0.768***	0.784***	0.690***	0.676***
	(0.026)	(0.025)	(0.076)	(0.075)	(0.097)	(0.098)	(0.181)	(0.180)
Bottom vs. Middle (p-value)	0.66							
Bottom vs. Top (p-value)	0.28							
Middle vs. Top (p-value)	0.13							
Top Proportion vs. Bottom Proportion (p-value)		0.24		0.87		0.13		0.71
Sample Size	1,200		370		580		250	

NOTES: All test scores are expressed in NCEs. Standard errors are clustered at the block level. All specifications control for block fixed effects. Student controls include gender and eligibility for free lunch. Teacher controls include teacher's type: AC or TC, gender, race/ethnicity and teaching experience. Average peer subject-specific achievement measured at the classroom level. The proportion of top 25% and bottom 25% of peers in a classroom are based on the grade and subject-specific baseline test score distribution. Due to confidential nature of the data, the sample sizes are rounded to the nearest tenth.
* significant at 10%, ** significant at 5%, *** significant at 1%.

Table 9: Tests of Peer Effect Models-Full Sample

Panel A: Endline Reading Test Scores	**Tests**			
Models (Number of Tests)	Number of Estimates Significant in Direction the Model Suggest	Number of Estimates Insignificant in Direction the Model Suggest	Number of Estimates Significant in Opposite Direction	Number of Estimates Insignificant in Opposite Direction
Weak Monotonicity (3)	0	0	1	2
Strong Monotonicity (6)	0	0	1	5
Invidious Comparison (4)	0	4	0	0
Ability Grouping (6)	2	1	0	3
Weak Frame of Reference (3)	1	2	0	0
Strong Frame of Reference (6)	1	5	0	0

Panel B: Endline Math Test Scores	**Tests**			
Models (Number of Tests)	Number of Estimates Significant in Direction the Model Suggest	Number of Estimates Insignificant in Direction the Model Suggest	Number of Estimates Significant in Opposite Direction	Number of Estimates Insignificant in Opposite Direction
Weak Monotonicity (3)	0	0	0	3
Strong Monotonicity (6)	0	2	2	2
Invidious Comparison (4)	2	0	0	2
Ability Grouping (6)	0	2	1	3
Weak Frame of Reference (3)	0	3	0	0
Strong Frame of Reference (6)	2	2	0	2

NOTES: See text for further details on the models and the tests conducted. The tests are based on the coefficients on the proportion of peers in the top 25% and bottom 25% from the full sample when own student varies by grade and subject specific placement in the baseline distribution.

Table 10: Estimates of Peer Effects by Identification Strategy and Peer Aggregation for the Repeated Cross-Section Sample

Dependent Variable: Endline Test Scores			
	Randomization Specification (Class Level Peer Achievement)	Repeated Cross-Section Specification (Class Level Peer Achievement)	Repeated Cross-Section Specification (Grade Level Peer Achievement)
Panel A: Reading Test Scores	(1)	(2)	(3)
Average Peer Baseline Achievement	-0.437***	-0.398***	-0.235
	(0.114)	(0.100)	(0.155)
Own Baseline Test Score	0.635***	0.638***	0.645***
	(0.039)	(0.045)	(0.045)
Sample Size		760	
Panel B: Math Test Scores	(3)	(4)	(5)
Average Peer Baseline Achievement	-0.277	-0.221**	-0.158*
	(0.204)	(0.103)	(0.082)
Own Baseline Test Score	0.601***	0.606***	0.606***
	(0.036)	(0.042)	(0.042)
Sample Size		710	

NOTES: All test scores are expressed in NCEs. Standard errors are clustered at the block level. The repeated cross-section sample includes the 10 schools that were present in both 2004-2005 and 2005-2006. Column (1) controls for block fixed effects and Columns (2) and (3) control for grade, school, and grade by year fixed effects. Student controls include gender, race/ethnicity and eligibility for free lunch. Teacher controls include teacher's type: AC or TC, gender, race/ethnicity and teaching experience. Average peer subject-specific achievement measured at the classroom level in columns (1) and (2) and the grade level in column (3). Due to the confidential nature of the data, the sample sizes are rounded to the nearest tenth.
* significant at 10%, ** significant at 5%, *** significant at 1%.

Appendix:

Table A1: Teacher Summary Statistics

	All Teachers	**AC Teachers**	**TC Teachers**
	Mean (Standard Error)	Mean (Standard Error)	Mean (Standard Error)
Female	0.90	0.87	0.93
	(0.29)	(0.33)	(0.25)
Race			
White	0.59	0.45	0.72
	(0.49)	(0.50)	(0.44)
Black	0.24	0.35	0.12
	(0.43)	(0.48)	(0.33)
Hispanic	0.17	0.20	0.15
	(0.38)	(0.40)	(0.36)
Experience	3.29	3.10	3.36
	(1.59)	(1.59)	(1.59)
Hours of Instruction for Certification			
Total	462.06	296.91	623.69
	(253.57)	(150.76)	(228.64)
Reading	89.28	61.32	116.65
	(57.31)	(44.95)	(55.02)
Math	33.15	25.44	40.85
	(23.18)	(23.14)	(20.61)
SAT (or SAT Equivalent) Composite Score	972.40	960.76	982.91
	(161.98)	(179.35)	(145.01)
# of Teachers	180	90	90

NOTES: Due to confidential nature of the data, the sample sizes are rounded to the nearest tenth.

Table A2: Tests of Peer Effect Models-Subgroups

Panel A: Endline Reading Test Scores		**Tests**		
Models (Number of Tests)	Number of Estimates Significant in Direction the Model Suggest	Number of Estimates Insignificant in Direction the Model Suggest	Number of Estimates Significant in Opposite Direction	Number of Estimates Insignificant in Opposite Direction
Weak Monotonicity (15)	0	0	2	13
Strong Monotonicity (30)	0	5	5	20
Invidious Comparison (20)	0	16	0	4
Ability Grouping (30)	7	5	0	18
Weak Frame of Reference (15)	2	13	0	0
Strong Frame of Reference (30)	5	20	0	5

Panel B: Endline Math Test Scores		**Tests**		
Models (Number of Tests)	Number of Estimates Significant in Direction the Model Suggest	Number of Estimates Insignificant in Direction the Model Suggest	Number of Estimates Significant in Opposite Direction	Number of Estimates Insignificant in Opposite Direction
Weak Monotonicity (15)	0	6	0	9
Strong Monotonicity (30)	0	12	4	14
Invidious Comparison (20)	3	7	0	10
Ability Grouping (30)	1	13	2	14
Weak Frame of Reference (15)	0	9	0	6
Strong Frame of Reference (30)	4	14	0	12

NOTES: See text for further details on the models and the tests conducted. The tests are based on the coefficients on the proportion of peers in the top 25% and bottom 25% from the subgroups (female students, male students, free-lunch eligible students, black students, and Hispanic students) when own student varies by grade and subject specific placement in pre-treatment distribution.

Table A3: Tests of Peer Effect Models-Full Sample with Bonferroni Corrections

Panel A: Endline Reading Test Scores	**Tests**		
Models (Number of Tests)	Number of Estimates Significant in Direction the Model Suggest	Number of Estimates Significant in Opposite Direction	Number of Insignifcant Estimates
Weak Monotonicity (3)	0	1	2
Strong Monotonicity (6)	0	1	5
Invidious Comparison (4)	0	0	4
Ability Grouping (6)	2	0	4
Weak Frame of Reference (3)	1	0	2
Strong Frame of Reference (6)	1	0	5

Panel B: Endline Math Test Scores	**Tests**		
Models (Number of Tests)	Number of Estimates Significant in Direction the Model Suggest	Number of Estimates Significant in Opposite Direction	Number of Insignifcant Estimates
Weak Monotonicity (3)	0	0	3
Strong Monotonicity (6)	0	0	6
Invidious Comparison (4)	0	0	4
Ability Grouping (6)	0	0	6
Weak Frame of Reference (3)	0	0	3
Strong Frame of Reference (6)	0	0	6

NOTES: See text for further details on the models and the tests conducted. The tests are based on the coefficients on the proportion of peers in the top 25% and bottom 25% from the full sample when own student varies by grade and subject specific placement in the baseline distribution.