Academic Differentiation, Classroom Peer Skill, and Inequality: Evidence from a Natural Experiment in 60 Urban High Schools

Takako Nomi

St. Louis University

Stephen W. Raudenbush

University of Chicago

This Draft: February 17, 2013

Key Words: Tracking, High School Mathematics, Classroom Peer Composition, Regression Discontinuity Design.

Research reported here was supported by a grant from the Institute of Education of the US Department of Edication. The grant is entitled “Making a Success of “Algebra for All”: A Fleet of Natural Experiments in Urban Curricular Reform.”

Abstract

Most US high school students attend non-selective comprehensive public schools. Within these schools, students have historically been assigned to mathematics classes on the basis of prior math skill. Critics argue that this practice, often referred to as "tracking," restricts learning opportunities for low-skill students, increasing social and ethnic inequality. In response, educators have launched several reforms: "de-tracking" high schools, requiring academic courses for all students, and providing extended instructional time to low-skill students. Studies of tracking and reform, however, have not isolated the impact of classroom peers' skill on student achievement. In this study we exploit a natural experiment conducted within each of 60 Chicago high schools in which skill-based segregation in mathematics was intensified while instructional time and content coverage were increased for low-skill students. Using a regression discontinuity design with school-specific instrumental variables, we find that assignment to a low-ability class sharply reduced mathematics achievement of typical Chicago students. Holding constant classroom peer ability, increasing instructional time substantially increased math achievement. These results reveal ways in which high-school academic organization shapes the distribution of human capital.

BACKGROUND AND SIGNIFICANCE

An enduring controversy in sociology concerns the role of schooling in the intergenerational transmission of inequality. Educational attainment varies by race, ethnicity, and social background and predicts adult employment, occupational status, earnings, and health (Hansen, Heckman, and Mullan, 2004; Herd, Pamela, Brian Goesling, and House, 2007 Raudenbush and Kasim, 1998). Controversy focuses on practices schools use to manage heterogeneity in student skills and whether these practices reinforce or moderate social and ethnic disparities in outcomes.

Academic Differentiation in US Schools

All modern societies assign children to classrooms based on age, yet even within same-age classrooms, children start the school year with substantially varied cognitive skills. In the US, the math skills of students entering kindergarten vary as a function of student socio-economic status and ethnicity (Lee and Burkham, 2002). Organizational devices to reduce skill heterogeneity within instructional units are pervasive if not universal. In most US elementary schools, children who make slow progress in academic skills can be retained in grade, and we find that low-SES and ethnic minority children are disproportionately retained (Jacob and Lefgren, 2004; Shepard, 1989; Hong and Raudenbush, 2005), or placed in special education (Harry and Klingner, 2005). Such a practice reduces skill heterogeneity while increasing age heterogeneity within classrooms. More pervasively, elementary school teachers routinely sub-divide children within their classrooms into homogeneous ability groups, particularly for reading instruction but often for mathematics instruction as well (Barr and Dreeben, 1983; Gamoran, 1986; Hallinan and Sorensen, 1983; Hong and Hong, 2009; Nomi, 2010; Pallas, Entwisle, Alexander, and Stluka, 1994; Slavin, 1987) .

Social and ethnic disparities in math and reading achievement grow slowly during the elementary years (Fryer and Levitt, 2004; Downey, Hippel, and Broh, 2004), and these differences are quite pronounced when children enter secondary schools. The secondary sector copes with the resulting skill heterogeneity in one of two ways: by selective admission to exam schools, private schools, and special magnet schools; and by differentiating coursework (tracking) within non-selective secondary schools (Kerckhoff, 1986; Lucas, 1999; Gamoran, 2010). Attending a selective school or taking "high-track" courses within a non-selective school is quite strongly associated with ethnic and social background and strongly predict access to post-secondary education and later adult outcomes (Alexander and Cook, 1982; Rosenbaum, 1976; Oakes, 1995, 2005; Vanfossen, Jones, and Spade, 1987).

Let us denote the organizational process that manages skill heterogeneity from kindergarten through post-secondary schooling as "academic differentiation." Critics argue that academic differentiation increases inequality in cognitive skills (Lucas, 1999; Rosenbaum, 1976; Oakes, 2005). Segregation of struggling students with low-achieving peers and the restriction of course content in low-ability settings may depress opportunities to learn (Gamoran, 1987, Gamoran and Mare, 1989). Continuity over time in placement in low-ability groups, tracks, or schools (Alexander and Cook, 1982; Gamoran, 1992; Hallinan, 1991) may create a cascade of negative effects as diminished opportunities one year lead to grouping or tracking decisions the next that further reduce opportunities, generating depressed educational attainment as adults. Because parental education is linked to pre-school home learning opportunities, academic differentiation then arguably contributes to the inter-generational transmission of inequality.

A contradictory view asserts that creating homogeneous instructional settings increases the efficiency of instruction. Teachers can target their instruction to the "zone of proximal development" of their students, challenging students to move to the next level of skill but not overwhelming them with unattainable learning goals (see review by Hallinan, 1994; Gamoran, 2010). Moreover, if children form academic self efficacy by comparing their own ability to those of classroom peers, it may be that assigning low-achieving students to classrooms that contain much higher-achieving peers may actually reduce the learning of the low achievers.

A longstanding related argument is that academic differentiation promotes social efficiency; as students reveal their skills and interests, the system of academic differentiation guides them to occupational roles for which they are best suited (Davis and Moore, 1945 Powell, Farrar, and Cohen, 1985). Resources for post-secondary education are concentrated on those who can benefit most from them. Parsons (1950) famously portrayed a key function of schools to be teaching children how to function in adult society. As children move from home to school, they encounter teachers who publicly evaluate the varied skills of same age peers. Children thereby become increasingly aware of their skills and interests are thus incrementally prepared psychologically to accept differentiated adult roles as they progress through school.

The Role of Mathematics in Academic Differentiation of US High Schools

The vast majority of US students attend comprehensive public high schools as opposed to private schools or selective high schools. These comprehensive high schools are heterogeneous with respect to students' incoming academic background. For this majority of students, academic differentiation at the high school level operates by tracking – that is, by differentiated course assignments within high schools. Through much of the 20th century, comprehensive public high school tracks were global: students participated in an academic track, a general track, or a vocational track. Those in the academic track took college preparatory classes in all academic subjects (math, English, history, the sciences, and foreign language). Course work in all subjects within the other two tracks – general and vocational – was intended to prepare students to move into the labor market after high school graduation (Lucas, 1999). Global tracking completely confounded two social processes: the segregation of classrooms on the basis of prior academic skills, which have been correlated with social class origins and ethnicity; and the provision of completely separate streams of course content to the students so segregated.

Over the past 25 years a more complex and nuanced system of differentiated course assignments has replaced the global tracking system. In principle, a student within a comprehensive public high school may take more advanced courses in English than in math, more advanced courses in foreign languages than in history, etc. Indeed, in many schools, assignment to courses in social studies and English is not based on prior achievement in these subjects (Delaney, 1991; Lucas, 1999).

However, mathematics course work has remained highly differentiated within most US high schools (Brunello and Checchi, 2007; Rose and Betts, 2004; Raudenbush, Rowan, and Cheong, 1993). The standard college-preparatory sequence of pre-algebra, algebra, geometry, more advanced algebra, pre-calculus (including analytic geometry and trigonometry) and finally calculus remains largely in place. In large public comprehensive high schools, there are more or less advanced algebra courses, geometry courses, etc., so that even within the college preparatory sequence, classes in these schools will be segregated on the basis of prior student achievement. Moreover, in most high schools, starting algebra late (e.g., in 10^{th} grade) is often regarded as an indicator of poor mathematics preparation, and students with lower math achievement will tend to be assigned to take algebra in 10^{th} grade rather than in 8^{th} or 9^{th} grade.

Although few contemporary US high schools practice global tracking officially, the sharp differentiation of coursework within mathematics tends to constrain class scheduling in such a way as to increase academic differentiation in all subjects. Students scheduled to take math class together will tend also to take other classes together. By the first year of high school (grade 9), math achievement is highly correlated with reading and science achievement, so non-math courses will tend to become somewhat segregated on the basis of prior achievement in these other domains (Lucas, 1999). Moreover, the sequence of academic science courses tends to parallel that of math, such that students aiming to go to college will often take biology, then chemistry, then physics, where chemistry requires more math than does high school biology, with physics requiring even more math than chemistry. Even English classes tend to be tracked on the basis of students' prior math achievement, inducing academic differentiation in prior English achievement, which is correlated with prior math achievement (Lucas, 1999). In a sense, then, mathematics "carries the water" for academic differentiation within contemporary US high schools (Lucas, 1999). Like the global tracking system of the past, academic differentiation based on differentiated course assignments has continued to confound classroom peer composition with the subject matter covered in high- versus low-skill classrooms.

A number of school districts have recently enacted reforms that aim to equalize course taking opportunities for low- and high-skill math students. One set of reforms (reviewed below) "de-tracks" schools, simultaneously reducing classroom skill segregation while exposing all students to a common stream of academic math courses starting with algebra. A second kind of reform also attempts to equalize course content, but provides extended instructional time to low-skill students. Unanswered in studies of these reforms is the question of whether assignment to classes of high-versus low-skill students influences math learning. As a result, we cannot discern how three aspects of high school academic organization --- course assignment, instructional time, and classroom peer skill --- combine to shape the distribution of human capital, and in particular, mathematics achievement.

In this study, we exploit a natural experiment replicated within each of 60 neighborhood high schools in Chicago. The aim was to insure that all students, including low-skill students, would be exposed to college-preparatory algebra in grade nine, while providing extra instructional time for low-skill students. Perhaps ironically, introduction of this reform induced a sharp increase in classroom segregation based on prior skill. However, the extent to which the reform induced such segregation varied quite dramatically from school to school, creating an opportunity to gauge the importance of classroom peer skill as a determinant of student learning. A key strength of the study is that its regression-discontinuity design affords a strong basis of causal inference, which has been elusive in most of the past research.

HIGH SCHOOL MATH TRACKING AND SOCIAL STRATIFICATION: THE EVIDENCE

Assignment to classrooms based on observable differences in prior math skill generates SES and ethnic differences in classroom peer composition, course pacing, and course content (Gamoran, 2010). This finding is essentially undisputed in the US empirical literature, which shows invariably that SES and majority race predict placement in higher-track classes; that these differences are mostly explainable by prior math skill; and that “high-track classes” cover more topics at a higher level than do lower-track classes.

Opponents of tracking cite an extensive literature asserting that high-school tracking increases inequality by suppressing the learning low-track students, who are disproportionately from of lower-SES and ethnic minority backgrounds (Oakes, 1985). However, conceptual and methodological challenges leave the validity of this argument in doubt.

Studies of Track Placement

Most studies that are critical of tracking for reinforcing achievement inequalities are studies of the impact of track placement. The causal question in these studies is whether students assigned to a low track would fare better if assigned to a higher track. Critics assert that tracking simultaneously segregates classrooms by peers' prior ability and restricts coverage of course content of low-track classes. Two mechanisms are implicated. First, classroom segregation arguably induces negative peer effects as opportunities to learn from peers diminish and role models for effective achievement behavior disappear. Explanations based on such peer effects do not rely on hypotheses about the second mechanism: restricted access to course content, identified by critics as a key factor accounting for the negative impact of low-track placement (Oakes, 1985). Disentangling these two processes is crucial for understanding how alternative forms of academic differentiation may affect overall achievement and achievement inequality: schools would need to change both curricular structure (e.g., what courses to offer) and how to sort students into different classrooms when offering alternative systems of academic differentiation. Peer segregation and instructional orientation are typically confounded: Teachers have been found to gauge the conceptual level and pacing of their instruction to the current ability level of the students they teach (Barr and Dreeben, 1983; Gamoran, 2010; Raudenbush, Rowan, and Cheong, 1993). Beyond the problem of specifying the mechanisms by which academic differentiation influences inequality, studies of tracking also face serious challenges to valid causal inference.

Causal inferences about the impact of track assignment have been elusive. Students are not randomly assigned to tracks; prior achievement and motivation, for example, are likely predictors of track placement. How can we remove the effects of selection bias in estimating the impact of track placement? We find three methodological approaches used in the empirical literature on track placement.

Conditioning on observed covariates using regression. The most common approach is to regress an achievement test score on an indicator of track placement, holding constant observed covariates within a linear regression model. The covariates typically include demographic background and prior academic achievement. Some studies also include prior course taking as predictors (e.g., Alexander and Cook, 1982; Vanfossen et al, 1987). Most studies using this approach find that, with no control for covariates, high track placement strongly predicts high achievement outcome. Controlling for covariates (Alexander and Cook, 1982; Gamoran, 1987) sharply reduces the coefficient for track placement, which nonetheless remains non-negligible and statistically significant. This regression-based approach requires the strong assumption: that there are no unmeasured confounders – that is, unobserved covariates that predict the potential outcomes of students under alternative track assignments controlling for the observed covariates. The problem is that researchers typically have access to a limited set of covariates. In contrast, who actually make the decisions about track placement – counselors, parents, or the students themselves – plausibly have more information than do the researchers. For example, students who have similar values on the observed covariates may vary in how motivated they are to pursue advanced math; this motivation may affect track placement and potential outcomes. If so, a regression analysis that omits a good measure of motivation will be biased.

An additional problem afflicts such regression-based studies of the impact of track placement: that of unwarranted extrapolation. The problem arises because students in different tracks have very different distributions of the observed covariates. This is known in the statistical literature as the problem of a lack of "common support" on pre-assignment covariates (Rubin, 1997). For example, students with very low prior math achievement may have a near-zero probability of being assigned to a high math track. Similarly, students with very high prior achievement have little probability of being assigned to a low track. In this setting, the data contain no information on how very low-achieving students would do if assigned to a high track because few comparable students would be found in the high math track. Nor could we find in the low track high-achieving students who are comparable to many students in the high track. For students who fall outside the region of common support, regression can nonetheless estimate an effect of track placement. Unfortunately, the information for such comparisons comes not from the data but rather from the assumption of linearity in the regression model: we use the linear model to "extrapolate" how a very low-achieving student would respond if assigned to the high track. Critically, a failure the linearity assumption can produce bias in the estimated impact of track placement, not only for students outside the region of common support, but for those inside the region of common support.

Moreover, the familiar strategy of controlling for school fixed effects may exacerbate rather than ameliorate this problem. Controlling for school fixed effects restricts comparisons to students attending the same school. Such students are presumably subject to a common but unobserved (to the researcher) set of rules or guidelines that determine track placement, a practice that would minimize overlap in covariates within schools between those place in higher versus lower tracks.

If the linearity assumption is wrong – and there is no way of checking this for the lowest or highest achievement students --- the results may be biased even if the set of covariates is very complete, i.e., there are no unobserved confounders. In sum, if students in different tracks differ substantially on their covariates, as seems likely, regression-based estimates of the impact of track placement will depend more on the investigator's model assumptions than on the data.

Matching on the propensity score. In response to the problem of the failure of common support some studies have matched students on their propensity to be placed in a particular track (Attewell & Domina, 2008). Using this approach, very low achieving students having near zero probability of being placed in the high group would be discarded, as would very high achieving students with no probability of falling into the low track. The analytic results will then depend on a particular "margin" of students: those who might plausibly be assigned to either of the tracks. While matching on the propensity score -- and hence discarding unmatchable cases – avoids extrapolation, it influences the research question: We are now interested in whether a certain margin of students would fare better or worse in a higher track. This is certainly an interesting question but we must not generalize findings about how the tracking system affects the potentially large number of students who are either very low or very high achieving. Perhaps more important, matching on the propensity score does not address the problem of unobserved confounding.

Attewell and Domina (2008) examined the effect of track placement among marginal students who might be placed in two adjacent track levels who are matched on their propensity scores. They found that students placed in a higher track out-performed observationally students placed in a lower track. However, the magnitude of this effect was smaller than that reported in regression-based studies.

Selection modeling and instrumental variables. To address the problem of no unobserved confounding, some investigators have adopted methods to work even in the presence of unobserved confounders. One approach is to rely on an instrumental variable – a variable that can affect the outcome only by influencing track placement. This approach requires three key assumptions (Angrist, Imbens, and Rubin, 1996). First, the instrumental variable must be "exogenous," meaning that, controlling for observed covariates, assignment to the instrument is effectively randomized: no unobserved confounders predict the value of the instrument within levels of the observed covariates. Second, the instrument must meet an "exclusion restriction:" its impact on the outcome must work solely through its impact on track placement. Third, the instrumental variable must fairly strongly predict track placement. A key challenge then is to find an instrument that satisfies critics that these assumptions have been met.

A related approach uses a selection model as developed by Heckman (1979). One allows the error term in the selection model (e.g., the model predicting track placement) to be correlated with the error term in the outcome model, thereby eliminating the assumption of no unobservable confounding. The selection model requires a valid instrumental variable for strong identification, however. Like the regression approach, the method of instrumental variables and the method of selection modeling also may be vulnerable to the failure of common support as described above.

Gamoran and Mare (1989) compared students in different tracks attending tracked schools. Also, Rose and Betts (2004) examined the effect of high school math course taking (e.g., vocational math, pre-algebra, algebra, geometry) on earnings ten years after graduation. These studies relied on school-level covariates as instruments, which include such variables as average initial math scores, school racial and SES composition, school course offering (e.g., percent of students in academic track and the number of advanced math courses), cohort size, and the region in which a school is located. Their findings showed that higher track placement, or enrolling in more advanced courses, lead to better students' outcomes. The choice of instrumental variables is open to dispute in such studies.

Almost all studies of track placement, regardless of their analytic methods, showed that students who are placed in higher tracks do better than students in lower tracks, and studies using the regression approach tend to produce the largest effects. These findings are often cited to support an argument that tracking would lead to greater achievement inequalities. An alternative set of studies assess this implication by comparing how students fare in a tracking regime rather than a no-tracking regime.

Studies of Tracking Versus No-tracking

Several researchers, using different analytic approaches, have compared students exposed to tracking versus similar students exposed to an alternative regime without tracking. Such studies can ask not only whether de-tracking improves achievement but also whether de-tracking is especially helpful to students who would have been placed in a low-track math class under a tracking regime. The problem of "common support" that undermines the credibility of studies of track placement is likely to be less salient in studies of tracking versus de-tracking.

Propensity score stratification. Hoffer (1992) developed a propensity score model for the track placement of students attending schools that use tracking. He then applied this model to students attending untracked schools. This enabled him to stratify all students according to their estimated probability of track placement if attending a tracked school. Within each stratum, this method removes observed confounding in assessing the impact of tracking on students who would hypothetically be exposed to varied track placement. He found that students assigned to a high track out-performed observationally similar students who were not tracked. This clever strategy assures common support on observed covariates but is, of course, vulnerable to unobserved confounding.[i] Moreover, the propensity-score based analysis would need to consider confounding at both student- and school-levels; students are selected into different tracks and schools are selected into the use of tracking.

Selection models and instrumental variables. Other studies comparing average achievement in tracked and untracked schools include Argys, Rees and Brewer (1996), who used Heckman's (1979) selection model, Betts and Shkolnik (2000), and Figlio and Page (2002), who used an instrumental variable approach. As the choice of instrumental variables, these authors typically rely on school-level characteristics, such as racial, SES, and academic composition, academic course offering, the region and urbanity, and political orientation in the county. The extent to which these instruments meet the key IV assumptions is open to criticism. Results, in general, showed no effect of tracking among students attending classes with relatively low average ability, while tracking may have a positive effect for students attending high-ability classrooms.

Interrupted time series. Another strand of research used an interrupted time series design by taking advantage of a sudden shift in tracking practices. Burris, Heubert, and Levin (2006) assessed the impact of a de-tracking reform on student outcomes in suburban schools which provided accelerate math instruction to all students, while additional help was provided to struggling students. The authors found that future math course taking and student achievement increased overall with no detriment to high-ability students. However, this study did not examine the extent to which classroom peer composition was affected by the detracking reform. Burris et al., (2006) note that the vast majority of students in their study schools had skills above the national average. In this context, the detracking reform might have had relatively small effect on classroom peer composition.

In contrast, Nomi (2012) evaluated the impact on high achievers of the "Algebra for All" in Chicago, a large urban district where more than half of its students enter high school with skills below the national average. Using an interrupted time-series design, she found that this reform effectively "de-tracked" Chicago high schools, equalizing algebra opportunities. For high achievers with skills above the sixty-five national percentile, their algebra enrollment was affected as they had always taken algebra, but their classroom peer ability levels declined considerably due to an influx of low-achieving students. As a result, their achievement was also negatively affected by the reform.

Summary

In sum, studies of track placement agree that students in higher track levels outperform students in lower tracks. The size of the effects appears to vary with the method of analysis, though all methods require strong assumptions for valid causal inference. In contrast, studies that compare similar students in tracked and untracked schools, addressing a question about tracking, produced varied results with some studies relying on somewhat weaker assumptions than others for causal identification. Critically, most studies confound the effects of course content and classroom peer composition. We conclude that considerable uncertainty remains concerning the role of academic differentiation, and classroom peer skill in math in the production of educational inequality.

THE CURRENT STUDY

In this study we exploit a natural experiment replicated within each of 60 non-selective Chicago neighborhood high schools. A policy shift implemented in 2003 enables us to study the causal effect of a sharp increase in classroom segregation based on prior peer achievement. Using a regression discontinuity design (Cook and Campbell, 1979 Cook, 2008), we can examine how this shift influenced the math achievement of typical students in these schools, and this causal inference is based on comparatively weak assumptions. Because of rather dramatic school-to-school variation in how the reform was implemented, we are able to disentangle the impact of classroom peer prior achievement from the effect of shifts in course content that exposed low-achieving students simultaneously to academic algebra and supplemental remedial math. A large sample size enables quite precise estimation of these effects. Because this study addresses the key conceptual and methodological challenges discussed above, we regard it as providing exceptionally important new evidence regarding the impact of segregating classrooms based on prior math achievement, thus defining a significant new contribution to the research on secondary school tracking and inequality. We caution that the regression discontinuity design restricts inferences to "typical" Chicago students, who are predominately minority and low income and whose 8th grade math achievement is near the middle of the city's distribution.

Background: High-school Curricular Reform in Chicago

Before 1997, some neighborhood public high schools in Chicago required all ninth graders to take algebra, but most schools emphasized remedial arithmetic for struggling ninth graders. Under a new "Algebra-for-All" policy inaugurated in 1997, all public high schools required ninth graders to take algebra. This reform immersed many low-achieving students, for the first time, in academic math coursework. Evidence suggests that Algebra for All was less successful in raising test scores than anticipated, plausibly because many students lacked sufficient mathematical background to benefit from instruction in algebra (Allensworth, Nomi, Montgomery, and Lee, 2010; Nomi, 2012). In response, Chicago school officials instituted a new citywide reform in 2003: "Double-Dose Algebra." This policy required all students scoring below a cut-point on the pre-test not only to take algebra but also to take an additional period of mathematics designed to enable those students to catch up on basic skills required to master algebra. This reform, therefore, doubled the math instructional time of low-achieving students while exposing them to a standard academic algebra class. Research suggests that this reform produced modest but highly statistically significant positive effects for those low-achieving students (Nomi and Allensworth, 2009).

We know that these two reforms changed more than the course content experienced by struggling math learners. Each reform also had remarkably large effects on the classroom composition of all ninth-grade math classes, including those experienced by high-achieving students. Algebra for All dramatically reduced classroom segregation based on prior math achievement, which we shall typically refer to as sorting or classroom sorting in this study. Not surprisingly, requiring all students to take courses previously taken only by high achievers generated an influx of low-achieving students into algebra classes. This influx made those algebra classes much more heterogeneous in prior math achievement than algebra classes had been in previous years. As a result, the policy simultaneously increased the rigor of course work for low-achieving students and *increased* the prior math achievement of their peers. Algebra for All was, in essence, a de-tracking policy.

Six years later, the Double-Dose Algebra policy, which is our focus here, substantially *increased* classroom sorting based on prior achievement. Scheduling constraints assured that, on average, students taking Double-Dose algebra would have *lower* ability peers than would students who scored above the cut point and who therefore took a single period of algebra instruction. In effect, Double-Dose induced segregation of high school math classes by student skill levels.

The dramatic effect of the "double-dose" policy on classroom segregation is apparent in Figure 1. The left panel of the figure displays classroom mean prior math achievement (vertical axis) as a function of student's own prior math achievement (horizontal axis) one year prior to the implementation of the Double-Dose policy. The slope of the line describing this association is an index of segregation: it is the fraction of variation in student achievement that is explained by classroom mean achievement. The positive slope of this line indicates that there is some degree of segregation, even before the policy the incoming math achievement test score (Iowa Test of Basic Skills or "ITBS") in academic year 2001-2002, one year prior to the implementation of the Double-Dose policy.

The right panel of the figure displays the same association, but now during the year after the implementation of the Double-Dose policy. We see a marked discontinuity at a pre-test score of 50. This is because all students scoring below 50 on the pre-test were expected to take Double Dose. Taking Double-Dose algebra sharply reduced the prior achievement of one's peers, generating, in principle, a powerful natural experiment that enables us to assess the impact of classroom peer ability for students in the neighborhood of the cut point. The gap in mean prior achievement between those just above the cut point and those just below the cut point can be regarded as the impact of the Double-Dose policy on classroom segregation based on prior achievement.

FIGURE 1 ABOUT HERE

Double-Dose Algebra and Classroom Peer Composition

A key complication, however, involves the association between classroom peer ability and course content. Those scoring below the cut point not only experienced substantially different peers from those who score above; they also tended to experience different course work, just as the policy intended. In particular, the policy stated that those scoring below the cut point should take two periods of mathematics: a regular algebra class supplemented by an extra hour of support coursework designed to help low-achieving students "catch up" in their math skills and thereby to achieve well in mathematics. Figure 2 shows that compliance with this policy, while far from perfect, was substantial.

FIGURE 2 ABOUT HERE

The horizontal axis is identical to that in Figure 1 while the vertical axis gives, for each possible pre-test score, the fraction of students who took Double-Dose Algebra. We see a sharp discontinuity at pretest score = 50: those scoring below are far more likely than those scoring above to take two periods of math.

Figures 1 and 2 would seem to imply that we cannot separate the impact of the policy shift on classroom peer ability from its effect on course taking. Yet, such a conclusion turns out to be incorrect. What enables us to disentangle these effects is the dramatic school-to-school variation in how the policy was implemented. Schools varied significantly in how fully they complied with Double Dose: Scoring below the cut point was significantly more predictive of taking two periods of math in some schools than in others. In the extreme were several schools that appeared not to implement the policy at all. In addition, schools varied even more in the extent to which taking Double Dose was associated with a shift in classroom peer ability.

The remarkable heterogeneity in how schools implemented the policy is displayed in Figure 3. The horizontal axis displays, for each of 60 high schools in our study, the difference between two proportions: the proportion taking Double Dose among those scoring below the cut point and the proportion taking double dose among those scoring above the cut point. This can be regarded as a measure of the extent to which schools complied with the policy. For most schools, compliance was high – in the neighborhood of .7 to 1.0. However, compliance was lower than .7 for a non-negligible number of schools. The vertical axis shows a difference between two means: the mean prior achievement of classroom peers of those students scoring just below the cut point and the mean achievement of classroom peers of those scoring just above the cut point. The vast majority of these mean differences are negative, an indication that scoring below the cut point reduced one's classroom peer ability. However, the mean differences vary substantially. In a non-trivial number of schools, implementation of the policy was associated with very little change in classroom peer segregation based on prior ability.

Importantly, the two dimensions of Figure 3 are only modestly correlated: Compliance on Double Dose (horizontal axis) is correlated by r = -.27 with the contrast in classroom mean prior achievement. The fact that the two dimensions vary but are only modestly correlated provides the basis for an analysis that can, under assumptions we shall discuss in detail, separate the impact of classroom peer ability from the impact of course taking for those students scoring around the cut point.

FIGURUE 3 ABOUT HERE

IDENTIFICATION STRATEGY

According to our conceptual model (Figure 4) scoring below the cut point simultaneously increases the probability that a student will take Double Dose Algebra and reduces the average prior achievement of that student's peers. These changes in course-taking and peer composition then influence the algebra test score. We assume that scoring below the cut point can affect the outcome only through one of the two specified mechanisms: changing course taking or changing mean peer ability.

FIGURE 4 ABOUT HERE

We develop a two-level, two-stage least squares model (Raudenbush, Reardon, and Nomi, 2012) to represent the relations in Figure 4. We specify the effect of scoring below the cut point and each of our causal variables of interest using a pair of first-stage equations. Within each school, we specify the model

$$\begin{aligned}(PEER_{ij}) &= \gamma_{0j} + \gamma_{1j}(X_{ij} - \overline{X}_j) + \gamma_{2j}[(X_{ij} - \overline{X}_j)^2 - S_{xj}^2] + \gamma_{3j}(C_{ij} - \overline{C}_j) + \varepsilon_{ij} \\ (DD_{ij}) &= \alpha_{0j} + \alpha_{1j}(X_{ij} - \overline{X}_j) + \alpha_{2j}[(X_{ij} - \overline{X}_j)^2 - S_{xj}^2] + \alpha_{3j}(C_{ij} - \overline{C}_j) + e_{ij}\end{aligned} \quad (1)$$

where our causal variables are $PEER_{ij}$, the mean prior achievement of the classroom peers of student *i* in school *j*, and $DD_{ij} = 1$ if student *i* in school *j* scores below the cut point, 0 otherwise; our instrumental variable C_{ij} =1 if that student scored below the cut point on the pretest, 0 if not; and X_{ij} is the prior math achievement score. We center the student-level predictors C_{ij}, X_{ij}, and

X_{ij}^2 around their school means; S_{xj}^2 is the school mean of $(X_{ij} - \overline{X}_j)^2$. School-mean centering removes all between-school confounding from this within-school model. Equation 1 is a generalization of the standard school fixed effects model that allows randomly varying coefficients (Raudenbush, 2009).[ii] According to Equation (1) each of our causal variables is a quadratic function of the pre-test X_{ij} plus a step-function at the cut point (the 50th percentile on the ITBS score). The school-specific impacts of scoring below the cut point on course taking and peers are γ_{3j} and α_{3j}. We allow the regression coefficients in (1) to vary randomly between schools. We estimated a parallel model to assess the impact of scoring below the cut point on algebra achievement. This is an "intent-to-treat" estimate, where scoring below the cut point expresses the intention to treat.

If we had a single instrument (scoring below the cut point) and two endogenous causal variables (course taking and peer composition), we could be unable to identify the causal effects of interest. However, the replication of the natural experiment within 60 schools generates 60 instrumental variables having heterogeneous first-stage impacts α_{3j} and γ_{3j} for schools $j=1,\ldots,J$. Kling, Liebman, and Katz (2007) used a similar strategy in their analysis of the "Moving to Opportunity" experiment; Reardon and Raudenbush (in press) define the assumptions required to draw causal inferences based on this strategy.

Thus, at the second stage, we formulate the following model,

$$Y_{ij} = \delta_{0j} + \delta_{1j}(X_{ij} - \overline{X}_j) + \delta_{2j}[(X_{ij} - \overline{X}_j) \ - S_{xj}^2] + \theta_3(PEER_{ij}) + \delta_3(DD_{ij}) + v_{ij} \qquad (2)$$

Where $PEER_{ij}$ and DD_{ij} are classroom peer ability and double-dose enrollment by Equation 1, each centered about their school means.[iii]

The key assumptions are as follows: (1) *Scoring below the cut point cannot influence the outcome except by changing coursework and peers as measured here*. (2) *Scoring below the cut point is independent of unobserved confounders given observed covariates*. As is standard in regression discontinuity designs, we control for a polynomial function of the pre-test in estimating all coefficients. We tested the validity of our quadratic model by testing our models using pre-policy cohorts and find no evidence of spurious "causal effects". (3) *The instrument significantly predicts both participation in double dose and classroom peer composition*. As is manifest in Figures 1 and 2, this assumption is upheld. (4) As Reardon and Raudenbush (in press) show, we must also assume that *the association between our instrument and at least one of our endogenous variables varies from school to school*. Both do, as is also manifest in Figure 3. (5) The *impact of scoring below the cut point on double dose coursetaking cannot not perfectly correlated with the impact of scoring below the cut point on classroom peer composition*. We see from Figure 3 that this is assumption is met, and note r=-.27, so that the association between these two effects is modest in size. (6) *The impact of taking double-dose algebra on achievement must be independent of a) the impact of scoring below the cut point on taking double dose algebra and b) the impact of scoring below the cut point on peer classroom ability*; (7) Similarly, *the impact of classroom peer ability on achievement is independent of (a) and (b) above*.

Reardon and Raudenbush (in press) show that assumptions (1)-(7) are sufficient to define δ_3 and θ_3 as the average impact of taking double dose algebra on achievement and the impact of classroom peer ability on achievement for those whose course-taking is affected by scoring below the cut point. These are "complier average treatment effects" (Angrist, Imbens, and Rubin, 1996) defined only for those whose prior math ability is in the neighborhood of the cut point, and whose course assignment depends on whether they score above or below the cut point. Reardon and Raudenbush (in press) show that inferences are comparatively robust to the failure of assumptions (6) or (7) when we have a strong instrument (i.e., scoring below the cut point strongly affects the probability of taking double dose and also strongly affects classroom peer ability) and when the impact of the instrument on the two causal variables (taking double dose and classroom peer ability) are not too highly correlated. Figures (1) and (2) suggest that the instrument is strong and Figure 3 suggests that the two causal variables are not highly correlated.

SAMPLE AND DATA

The Chicago Public Schools is the third-largest school system in the United States. The district serves predominantly low-income and minority students; approximately 85% of students are eligible for free/reduced-price lunch and racial composition is about 50% African-American, 38% Latino, 9% White, and 3% Asian and other races/ethnicities.

We used data on 11,296 first-time ninth graders attending 60 non-selective, comprehensive neighborhood public high schools in Chicago during the 2003-2004 academic year. We exclude students with disabilities, which composed of 18 percent of the first time ninth graders, and most of them had incoming math skills below the 20^{th} percentile. Many schools exempted students with disabilities from taking double-dose algebra, and these students often took regular algebra with other students with disabilities, which would differ from regular algebra classes that regular below-norm students would have taken had they scored above the cut point. We also excluded students without valid classroom data. Four schools are excluded because they did not offer double-dose algebra at all, or they put all students in double-dose algebra.

Of students included in our analysis, approximately 86% were eligible for free or reduced lunch; 55 percent were African American and 34 percent were Hispanic (See Table 1).[iv] Students scoring below the 50th percentile on the Iowa Test of Basic Skills during the spring of their 8th grade year were expected to take Double-Dose Algebra. In fact, 82% of those scoring below the cut point did take Double Dose while 4% scoring above the cut point and who were therefore not expected to take Double Dose did in fact take Double Dose. Thus, compliance with policy, though not perfect, was high as suggested by Figure 1. Our outcome variable is the algebra subtest of the PLAN math test, developed by the American College Testing Service and administered to all students during the fall of their 10th grade year. The subtest contains 22 items; raw scores were converted to a scale score, and this was standardized with a mean=0 and standard deviation of 1. The key covariate for our analysis is the percentile score on the grade 8 ITBS test; a quadratic function (see Equation 2) sufficed to represent adequately the relationship between this covariate and the outcome. To measure classroom peer achievement, we first created a latent math score, using a vector of Iowa Test of Basic Skills Math Test (ITBS) scores from third through eighth grade.[v] We then computed the median of a student's classmates on this latent math score.

TABLE 1 ABOUT HERE

FINDINGS

First-Stage Results

We first consider the first stage of our analysis: To what extent does scoring below the cut-off point induce a shift in classroom peer ability and in the probability of taking Double Dose (that is, two periods of math)? We estimate models (1) and (2), each of which includes a school-specific quadratic function of the pre-test and school-specific step functions. Here scoring below the cutoff is a school-specific instrumental variable under our regression discontinuity design.

Both effects are strong, on average, but vary substantially and significantly from school to school (see Table 2). For students whose true prior achievement is around the cut point, scoring below the cut point induces a substantial reduction in classroom peer math achievement (see column 2): on average, scoring below the cut point induces a reduction of $\hat{\gamma}_3 = -0.24$, se = 0.03, about 0.35 standard deviations on the scale of the pre-test. However, this effect varies significantly and substantially from school to school, $\tau_\gamma^2 = 0.05, \chi^2(59) = 491.24, p = 0.00$. The large variance of this effect implies that in some schools scoring below the cut point has little or no effect on classroom peer achievement, while in other schools the effect is very large indeed.

To see this, let us compare our model's predictions for the modest effect of scoring below the cut point, that is, an effect that is one-standard deviation above average, to a school with a more extreme effect, that is one standard deviation below the average. These predictions are, respectively $\hat{\gamma}_{3j} = \gamma_3 + \sqrt{0.05} \approx -0.24 + 0.25 = .01$ and $\hat{\gamma}_{3j} = \gamma_3 - \sqrt{0.05} \approx -0.24 - 0.25 = -0.49$. The difference of 0.48 between these is a standardized effect size, given the standardization of the outcome variable (sd=1). The large variation across schools underscores our conclusion that we have here, in reality, 60 independent natural experiments in which scoring below the cut point induces varying degrees of change in classroom peer composition.

Table 2 also shows a large impact of scoring below the cut point on taking double dose (see column 3). Specifically, we estimate the average effect to be $\hat{\alpha}_3 = 0.72$, se = 0.03, implying that, for a student whose true prior achievement is near the cut point, scoring below the cut point increases the probability of taking double dose by about 72 percent, clearly large effect. However, an effect that also varies significantly and substantially from school to school, $\tau_{\alpha}^{2} = 0.04$, $\chi^{2}(59) = 552.73, p = 0.00.$ Finally, controlling for the pre-test, scoring below the cut point increase algebra achievement, on average, by a standardized effect size of 0.15 (se=0.05). This is the average intent-to-treat (ITT) effect, and this impact also varies significantly from school to school.

TABLE 2 ABOUT HERE

The first stage results are plotted in Figure 3. The estimated impact of scoring below the cut point on classroom peer achievement, that is $\hat{\gamma}_{3j}$, is displayed on the vertical axis while the estimated impact of scoring below the cut point on the probability of taking double dose, that is $\hat{\alpha}_{3j}$, is displayed on the horizontal axis. The correlation between these two effects is modest, $Corr(\hat{\gamma}_{3j}, \hat{\alpha}_{3}) \approx -.27.$

Second-Stage Results

Estimation of Equation 2 by regression via OLS or HLM would produce biased estimates of θ_3 and δ_3 if classroom assignment of peer composition or Double-Dose depends on covariates other than the pre-test, *X*. Such a regression estimate would also be subject to biases induced by a lack of common support on the pretest as described above. Instead, we adopt the instrumental variable method, replacing $(PEER_{ij})$ and (DD_{ij}) with predicted values $(\hat{PEER}_{ij})$ and $(\hat{DD}_{ij})$, generated by estimating the first-stage model.

Results (Table 3) indicate, most importantly, a strong positive impact of classroom peer ability, holding constant Double Dose, $\hat{\theta}_3 = 0.39$, se=0.13, t=3.04, p<.01. This implies that if students had experienced a rise in the average peer ability by one standard deviation, their algebra scores would improve by nearly 40 percent of a standard deviation, holding constant Double Dose. Controlling for classroom mean achievement, we find a large and significantly positive effect of taking double-dose algebra, $\hat{\delta}_3 = 0.31$, se=0.06, t=5.24, p<.001. This implies that had peer academic composition remained the same, taking double-dose algebra would improve algebra scores by about 0.31 standard deviations.

TABLE 3 ABOUT HERE

Specification Checks

We checked the specification of our model in several ways. First, ours is an additive model for the impact of classroom peer ability and the impact of taking double-dose algebra. The model also assumes an exclusion restriction – that the impact of scoring below the cut point operates solely through these two channels. If our model is correct, our estimates of the impacts of classroom peer ability and double dose should sum to the average ITT effect. We find this to be the case. Specifically, the double-dose algebra policy led to declines in peer ability for below norm students; scoring below the cut point resulted in lower classroom peer ability by 0.24 standard deviations (column 2 of Table 2). Thus, the average effect on algebra scores of 0.39 of scoring below the cut point (Table 3) mediated through peer ability change is estimated as (-0.24)*(0.39) =-0.094. Next, we note that, on average, scoring below the cut point increased double-dose algebra enrollment by 0.72 (column 3, Table 2); thus, the average policy effect on algebra scores via double-dose algebra, holding constant peer ability was (0.72)*(0.31)= 0.22. Putting these results together, model (3) predicts an ITT effect of 0.13. Our ITT analysis on algebra scores in (column 1 of Table 2) showed an effect of 0.15. Our model 3 thus reasonably reproduces this average ITT effect.

Second, our regression discontinuity analysis depends on the accuracy of the functional form in Equation 2, that is, the quadratic model for the association between pre and post test. To check this, we estimated our stage 1 model for the impact of scoring below the cut point on classroom composition – using data on a cohort of students who attended ninth grade one year prior to the implementation of the policy. We found no hint of a "below cut effect" on peer composition. We also conducted a similar pre-policy analysis of the impact of scoring below the cut point on algebra scores using the pre-policy cohort of ninth graders, again, finding no suggestion of an "effect." Withstanding these checks increases credence to the model, but of course does not insure that the model is correct.

DISCUSSION

School reforms that attempt to alter the structure of academic differentiation are likely to affect both access to specific coursework and how schools sort students into classrooms. This study highlights the importance of classroom peer academic composition in affecting student math learning independently of access to coursework. We showed that the effect of classroom peer ability levels is as consequential as specific coursework made available to students.

Our study can be regarded as a synthesis of 60 independent natural experiments induced by the city of Chicago's "Double-Dose Algebra" policy, inaugurated in 2003. The policy mandated that all students take academic algebra in ninth grade, but that students scoring below the mean on a pre-test to take two periods of ninth-grade math: one period of academic algebra and a second period of remedial math designed to support algebra learning of students who had fallen behind. On average, this policy tended to increase instructional time and, hence content coverage for low-achieving students while also inducing heightened between-classroom segregation based on prior math achievement. We found remarkable heterogeneity in how these 60 schools implemented the policy. In most but not all schools, compliance with assignment to double dose was reasonably high. However, the schools varied enormously in the extent to which implementation induced classroom peer segregation. This heterogeneity enabled us to disentangle the effects of two aspects of high school tracking: a) the impact of classroom peer skill; and b) the impact of access to course content via increased instructional time.

Controlling for instructional time, we find strong positive effects of classroom peer skill on math learning. If this relation holds for all students, we can anticipate that increasing classroom segregation based on prior achievement will tend to exacerbate inequality in algebra outcomes. (Such a finding does not imply, however, that achievement would increase overall.) Controlling for classroom peer skill, we find strong positive effects of increased instructional time. Taking an additional remedial class in addition to the regular algebra class substantially increases math learning.

We find that the most effective schools for typical children in Chicago neighborhood high schools are those that effectively encourage students scoring below the cut point to take double-dose algebra while at the same time minimizing classroom segregation based on prior achievement. It appears that careful attention to course scheduling can achieve this goal in many schools. This reasoning is based on further analyses that explored whether “mechanical factors” such as the mean and variance of a school's prior math achievement, strongly determine classroom peer composition. We found no evidence that this is the case.

A limitation of our design is that our findings apply only to those in the broad middle of the achievement distribution who might score either below or above the cut point. We estimate that about half of the students in our sample had a non-negligible probability of scoring either above or below the cut point. These are the students to whom our results apply. Roughly one quarter of the students had near zero probability of scoring above the cut point while a similar number had virtually no chance of scoring below the cut point. Our findings do not provide information on how these extremely low-scoring students (who have negligible risk of scoring above the cut point) or extremely high achieving students (who have negligible risk of scoring below the cut point) might respond to shifts in classroom peer achievement or extra instruction. To identify these effects will require an alternative research strategy and will constitute a topic for subsequent research.

In sum**,** our study cannot tell us whether de-tracked schools are more egalitarian than are tracked schools. However, our study does remove the confounding of classroom peer composition and instructional time, which is related to content coverage. For typical low-income minority students in our study, segregation into classes with low-achieving peers reduced achievement while increasing instructional time and hence, content coverage, increased achievement.

Clearly some students will have to attend classrooms composed of student who have low initial skill. Nomi (2012) suggests that doing so reduces the achievement of high-skill student. Our study suggests that doing so reduces the learning of a large group of students in the middle range of the Chicago math skill distribution. We currently have no strong evidence on how classroom peer skill affects the achievement of the lowest quarter of the math distribution. Nevertheless, these results indicate that how schools distribute the scarce resource of initial peer skill will shape the distribution of math skill. Instructional time in mathematics is also a scarce resource, but there may be significant scope for increasing it. What is required is an analysis of what students taking double dose would be doing if not taking an extra period of mathematics. Our study cannot prescribe how schools should organize peer skill and instructional time; but it does suggest that how schools do so will be consequential for the distribution of human capital as youth emerge from adolescence.

REFERENCES

Alexander, Karl and Martha Cook. 1982. "Curricula and Coursework: A Surprise Ending to a Familiar Story." *American Sociological Review* 47:626-640.

Allensworth, Elaine, Takako Nomi, Nicholas Montgomery, and Valerie E. Lee. 2009. "College Preparatory Curriculum for All: Academic Consequences of Requiring Algebra and English I for Ninth Graders in Chicago." *Education Evaluation and Policy Analysis* 31:367-391.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. " Identification of causal effects using instrumental variables." *Journal of the American Statistical Association* 91:444-472.

Argys, Laura M., Daniel I. Rees, and Dominic J. Brewer. 1996. "Detracking America's Schools: Equity at Zero Costs?" *Journal of Policy Analysis and Management* 15:623-645.

Attewell, Paul and Thurston Domina. 2008. "Raising the bar: Curricular intensity and academic performance." *Educational Evaluation and Policy Analysis* 30:51-71.

Barr, Rebecca and Robert Dreeben. 1983. *How Schools Work*. Chicago: University of Chicago Press.

Betts, Julian R. and Jamie L. Shkolnik. 2000. "The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools." *Economics of Education Review* 19:1-15.

Brunello, Giorgio and Daniele Checchi. 2007. "Does school tracking affect equality of opportunity? New international evidence." *Economic Policy* October.

Burris, Carol Corbett, Jay P. Heubert, and Henry M. Levin. 2006. "Accelerating mathematics achievement using heterogeneous grouping." *American Educational Research Journal* 43:105–136.

Cook, Thomas D. 2008. ""Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics." *Journal of Econometrics* 1422:636–54.

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin Co.

Davis, Kingsley and Wilbert E. Moore. 1945. "Some principles of stratification " *American Sociological Review* 10:242-249.

Delany, Brian. 1991. "Allocation, Choice, and Stratification within High Schools: How the Sorting Machine Copes." *American Journal of Education* 99:181-207

Downey, Douglas B., Paul von Hippel, and Beckett Broh. 2004. "Are Schools the Great Equalizer? Cognitive Inequality During the Summer Months and the School Year." *American Sociological Review* 69:613-635.

Figlio, David N. and Marianne E. Page. 2002. "School choice and the distributional effects of ability tracking: Does separation increase inequality? ." *Journal of Urban Economics* 51:497–514.

Fryer, Roland G. and Steven D. Levitt. 2004. "Understanding the Black-White Test Score Gap in the First Two Years of School." *The Review of Economics and Statistics* 86:447-464.

Gamoran, Adam. 1986. "Instructional and Institutional Effects of Ability Grouping." *Sociology of education* 59:185-198.

—. 1987. "The Stratification of High School Learning Opportunities." *Sociology of education* 60:135-155.

—. 1992. "Access to Excellence: Assignments to Honors English Classes in the Transition from Middle to High School " *educational evaluation and policy analysis* 14:185-204.

—. 2010. "Tracking and Inequality: New Directions for Research and Practice." Pp. 213-228 in *The Routledge International Handbook of the Sociology of Education*, edited by M. Apple, S. J. Ball, and L. A. Gandin. London: Routledge.

Gamoran, Adam and Robert D. Mare. 1989. "Secondary School Tracking and Educational Inequality: Compensation, Reinforcement, or Neutrality?" *American journal of sociology* 94:1146-1183.

Gamoran, Adam and Martin Nystrand. 1994. "Tracking, instruction and achievement " *International Journal of Educational Research* 21:217-231.

Hallinan, Maureen T. 1991. "School Differences in Tracking Structures and Track Assignments." *Journal of Research on Adolescence* 1:251-275.

—. 1994. "Tracking: From Theory to Practice." *Sociology of education* 67:79-97.

Hallinan, Maureen T. and Aage B. Sorensen. 1983. "The Formation and Stability of Instructional Groups." *American Sociological Review* 48:838-851.

Hansen, Karsten T., James J. Heckman, and Kathleen J. Mullan. 2004. "The Effect of Schooling and Ability on Achievement Test Scores." *Journal of Econometrics* 121:39-98.

Harry, Beth and Janette Klingner. 2005. *Why Are So Many Minority Students in Special Education? Understanding Race and Disability in Schools*. New York: Teachers College Press.

Herd, Pamela, Brian Goesling, and James S. House. 2007. "Socioeconomic Position and Health: The Differential Effects of Education versus Income on the Onset versus Progression of Health Problems." *Journal of Health and Social Behavior* 48:223-238.

Hoffer, Thomas B. 1992. "Middle School Ability Grouping and Student Achievement in Science and Mathematics." *educational evaluation and policy analysis* 14:205-227.

Hong, Guanglei and Yihua Hong. 2009. "Reading Instruction Time and Homogeneous Grouping in Kindergarten: An Application of Marginal Mean Weighting Through Stratification." *educational evaluation and policy analysis* 31:54-81

Hong, Guanglei and Stephen W. Raudenbush. 2005. "Effects of Kindergarten Retention Policy on Children's Cognitive Growth in Reading and Mathematics." *educational evaluation and policy analysis* 27:205-4224.

Jacob, Brian and Lars Lefgren. 2004. "Remedial education and student achievement: A regression-discontinuity analysis." *The Review of Economics and Statistics* 86:226-244.

Kerckhoff, Alan C. 1986. "Effects of Ability Grouping in British Secondary Schools." *American Sociological Review* 51:842-858.

Kling, Jeffrey R. , Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental analysis of neighborhood effects." *Econometrica* 75:83-119.

Lee, Valerie E. and David T. Burkam. 2002. " Inequality at the Starting Gate: Social Background Differences in Achievement as Children Begin School." Economic Policy Institute, Washington, D.C.

Lucas, Samuel R. 1999. *Tracking Inequality: Stratification and Mobility in American High School*. New York: Teachers College Press.

Nomi, Takako. 2010. "Educational stratification in early elementary school: The causal effect of ability grouping on reading achievement." *Journal of Research on Educational Effectiveness* 3:56-92.

—. 2012. "The Unintended Consequences of an Algebra-for-All policy on High-skill Students: the Effects on Instructional Organization and Students' Academic Outcomes." *Educational Evaluation and Policy Analysis* 34 489 - 505.

Nomi, Takako and Elaine Allensworth. 2009. "Double-Dose" Algebra as an Alternative Strategy to Remediation: Effects on Students' Academic Outcomes." *Journal of Research on Educational Effectiveness* 2:111-148.

Oakes, Jeannie 1985. *Keeping track: How schools structure inequality*. Birmingham, N.Y.: Vail-Ballou Press.

—. 2005. *Keeping track: How schools structure inequality* New Haven: Yale University Press.

Pallas, Aaron M., Doris R. Entwisle, Karl L. Alexander, and Francis M. Stluka. 1994. "Ability-Group Effects: Instructional, Social, or Institutional?" *Sociology of education* 67:27-46.

Parsons, Talcott. 1950. "The prospects of sociological theory." *American Sociological Review* 15:3-16.

Powell, Arthur G, Eleanor Farrar, and David K Cohen. 1985. *The shopping mall high school: Winners and losers in the educational marketplace.* . Boston: Houghton Mifflin.

Raudenbush, Stephen W. 2009. "Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings." *Journal of Education, Finance and Policy*. 4:468 – 491.

Raudenbush, Stephen W, Sean Reardon, and Takako Nomi. 2012. "Statistical analysis for multi-site trials using instrumental variables." *Journal of Research and Educational Effectiveness* 5:303-332.

Raudenbush, Stephen W, Brian Rowan, and Yuk Fai Cheong. 1993. "The pursuit of higher order instructional goals in secondary schools: Class, teacher, and school influences. ." *American Educational Research Journal* 30:523-553.

Raudenbush, Stephen W. and Rafa M. Kasim. 1998. "Cognitive skill and economic inequality: Findings from the National Adult Literacy Survey." *Harvard Educational Review* 68:33-80.

Reardon, Sean and Stephen W Raudenbush. Inpress. "Under what assumptions do site-by-treatment instruments identify average causal effects? ."

Rose, Heather and Julian R. Betts. 2004. "The effect of high school courses on earnings." *The Review of Economics and Statistics* 86:497-513.

Rosenbaum, James E. 1976. *Making Inequality: The Hidden Curriculum of High School Tracking*. New York: Wiley.

Rubin, Donald B. 1997. "Estimating causal effects from large data sets using propensity scores." *Annals of Internal Medicine* 127:757–763.

Shepard, Lorrie A. 1989. "A review of research on kindergarten retention." Pp. 64–78 in *Flunking grades: Research and policies on retention* edited by L. A. Shepard and M. L.Smith. London: The Falmer Press.

Slavin, Robert E. 1987. "Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis." *Review of educational research* 57:293-336.

Vanfossen, Beth E., James D. Jones, and Joan Z. Spade. 1987. "Curriculum Tracking and Status Maintenance." *Sociology of education* 60:104-122.

Notes

[i] However, to estimate the effect of track placement, he subdivided tracked students into three groups based on the observed track placement (high, middle, and low tracks) in each propensity-score stratum. He then estimated stratum-specific effect of track placement by comparing the outcome of each group to the average outcome of *all* untracked students. This comparison would produce biased estimates if unobserved covariates predict track placement and the outcome, given stratum membership.

[ii] We found no evidence that the two coefficients for the quadratic terms varied from school to school and hence set the between-school variance for these coefficients to zero.

[iii] As above, we found no evidence that coefficient for the quadratic terms varied from school to school and hence set the between-school variance for these coefficients to zero.

[iv] Data were made available by the Consortium for Chicago School Research (CCSR) at the University of Chicago with permission from the Chicago Public Schools (CPS). CCSR has a data-sharing agreement with CPS that enables it to complete administrative records on all students for each semester since 1991.

[v] Students' ITBS scores were first equated through Rasch analysis to remove form and level effects. Then, a two-level HLM, nesting years within students, modeled each student's learning trajectory; level 1 included variables for grade and grade-squared which were allowed to vary across students. There was also a dummy variable representing a repeated year in the same grade, to adjust for learning that occurred the second time in a grade

TABLES AND FIGURES

Figure 1. Classroom average skill levels by math percentile scores.

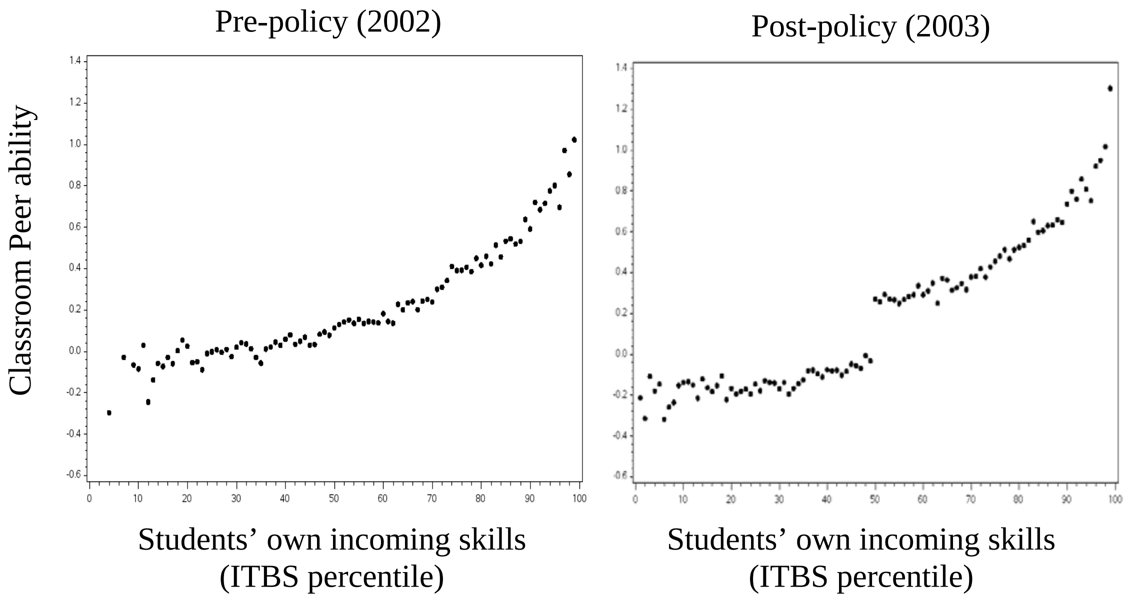

Figure 2. Double-dose Algebra Enrollment Rate by Math Percentile Scores

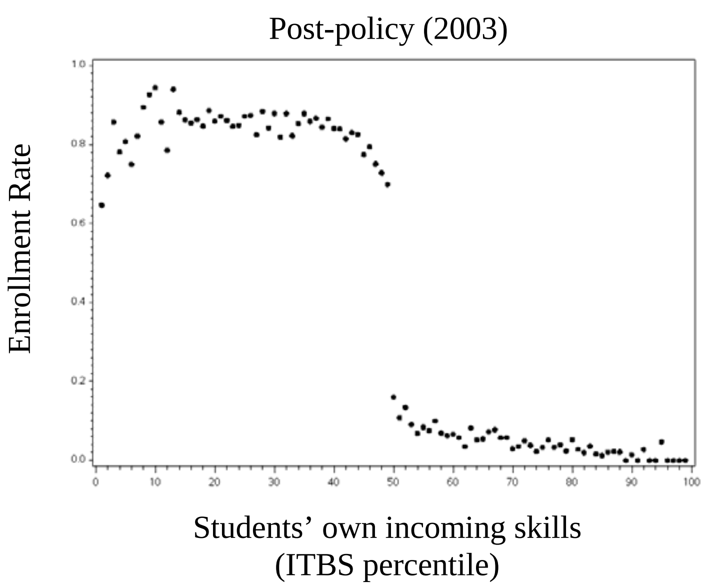

Figure 3. The effect of cut-score on Double-dose Algebra enrollment by the effect of cut-score on classroom peer ability

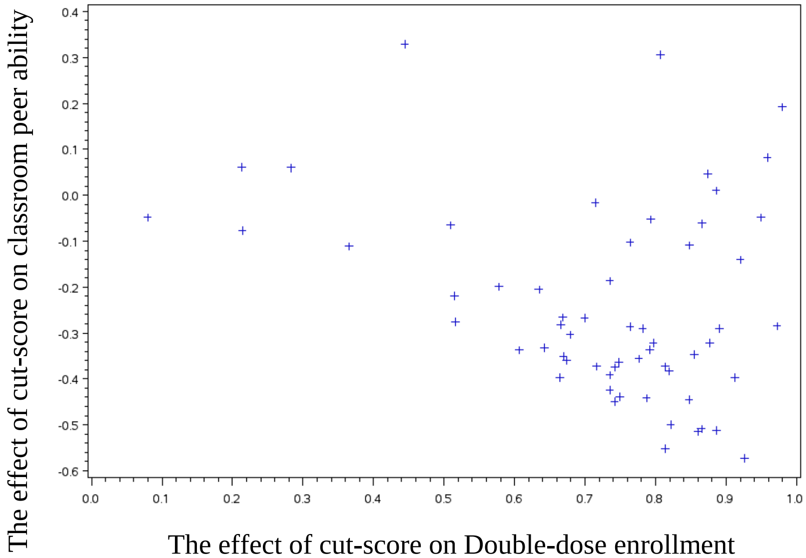

Figure 4. Causal model

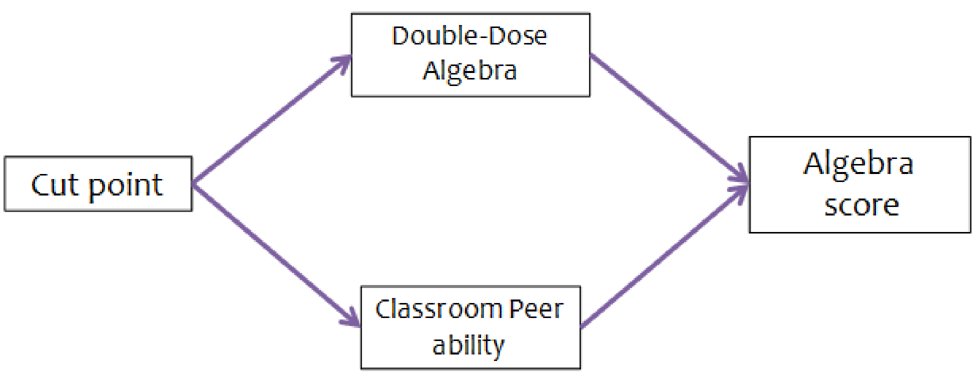

Table 1. Descriptive Statistics

Variables	Mean	SD
Scoring below cut point	0.40	0.49
Latent math skills	0.20	0.70
Algebra scores	0	1
Taking double-dose algebra	0.35	0.48
Class peer ability	0.18	0.58
Free/reduced lunch	0.86	0.35
African Americans	0.55	0.50
White	0.07	0.26
Hispanic	0.34	0.48
Asian	0.03	0.18

N=11,296

Table 2. Effect of cut point on double-dose algebra enrollment and classroom academic composition

	Algebra scores		Classroom academic composition		Double-dose algebra enrollment	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Intercept	-0.10	0.05	0.08	0.04	0.41	0.03
Below cut	0.15	0.03	-0.24	0.03	0.72	0.03
ITBS	0.024	0.0014	0.01	0.00099	-0.00014	0.00045
$ITBS^2$	0.00024	0.00002	0.00009	0.00002	-0.00002	0.00002
	Variance component	Chi-sq df=59	Variance component	Chi-sq df=59	Variance component	Chi-sq df=59
Intercept	0.15	3729.00	0.11	16218.91	0.04	6349.37
Below cut	0.03	89.08	0.05	491.24	0.04	552.73
ITBS	0.00011	496.52	0.00006	1642.54	0.00001	226.07
Level-1 σ^2	0.55		0.09		0.07	

Table 3. Estimated Effect of classroom peer ability, controlling for double-dose algebra and effect of double-dose algebra, controlling for classroom peer ability on algebra scores

	Algebra scores	
	Coefficient	SE
Intercept	-0.10	0.05
Double dose	0.31	0.06
Classroom peer ability	0.39	0.13
ITBS	0.02	0.0013
$ITBS^2$	0.00021	0.00002
	Variance component	Chi-sq df=59
Intercept	0.15	3746.57
ITBS	0.00005	411.31
$ITBS^2$	$1.4*10^{-8}$	152.66
Level-1 σ^2	0.55	3746.57