Making a Success of "Algebra for All": The Impact of Extended Instructional Time and Classroom Peer Skill in Chicago

Takako Nomi
Saint Louis University
Stephen W. Raudenbush
University of Chicago

In 2003, Chicago launched "Double-Dose Algebra," requiring students with pretest scores below the national median to take two periods of math–algebra and supplemental coursework. In many schools, assignment to Double Dose changed the peer composition of the algebra classroom. Using school-specific instrumental variables within a regression-discontinuity design (RDD), we find that attending a lower skill classroom reduced math achievement for median-skill students. As a result, the Double-Dose policy had little or no effect for median-skill students in schools that exposed them to low-skill classrooms. However, the effects of Double Dose were substantially positive in schools that did not do so. We consider policy implications and interpretations of the results from RDDs.

CHICAGO PUBLIC SCHOOLS (CPS) officials inaugurated a new citywide reform, "Double-Dose Algebra," in 2003. This policy required all ninth-grade students scoring below the 50th percentile on the eighth-grade math test to take two math classes: a regular algebra class plus a remedial class designed to help those students to catch up on skills required to master algebra. For those scoring above the 50th percentile, the policy continued to require them to take regular, "Single-Dose" algebra without an additional remedial class. Thus, the reform doubled the math instructional time of most low-achieving students attending neighborhood high schools in Chicago.

However, as we demonstrate in the current article, the "Double-Dose Algebra" policy also substantially *increased* classroom sorting based on prior math skill, on average, across Chicago's 60 neighborhood high schools. As a result, for many students whose prior skill was near the median, assignment to Double Dose entailed attending an algebra class composed of comparatively low-skill peers whereas assignment to single dose entailed attending a class of comparatively high-skill peers. This suggests that, for many median-skill students, "Double-Dose Algebra" is not a single treatment, but rather a mix of two concurrent treatments: doubled instructional time and reduced classroom peer skill. An important question is whether taking a class with comparatively low-skill peers enhanced or undermined the Double-Dose reform for students of median prior skill.

If the reform had increased sorting to the same extent in every school, we would have no basis to answer this question. However, this was not the case: In some schools, students assigned to Double Dose were concentrated in classrooms characterized by low prior math skill; in contrast, other schools managed to implement the policy without resorting to skill-based sorting, such that

classroom peer skill was relatively similar in Double-Dose and Single-Dose classrooms.

To study the impact of extended instructional time and peer math skill on algebra achievement, we exploit a "fleet of natural experiments," one occurring within each of 60 neighborhood public high schools. The large heterogeneity in how schools implemented the policy enables us to employ an instrumental variable strategy that, under "Assumptions" section we discuss in detail, identifies the impact of classroom instructional time and peer math skill on students of average prior math skill.

We conclude that the Double-Dose policy was modestly effective, on average, for median-skill students attending these 60 neighborhood high schools that serve most of Chicago's low-income students. However, the reform was substantially effective for these students only in schools where the policy did not induce large shifts in classroom peer math skill. The large policy impacts in those schools reveal the potential promise of the Double-Dose reform, a promise that is obscured by considering only the average impact across all schools. A caveat is that our results apply only to students whose prior achievement is in the neighborhood of the national median on pretest. However, because the pretest was measured with error, we also find that the fraction of students to whom our results apply is surprisingly large. Nevertheless, to understand how these policies may affect the entire achievement distribution, more needs to be known about how increased instructional time and classroom peer skill affect very low and very high achievers.

In the next section, we consider the relevance of this study broadly with respect to inequality in math achievement in the United States and more narrowly for curricular reforms now under way in many school districts. We review research investigating how assignment to low-skill classrooms may affect learning and, based on this literature, develop our hypotheses. We then describe our data and measures, our theoretical model, statistical methods and assumptions, findings, and implications for policy and program evaluation and reform. Last, we consider broad implications of our experience for conceiving and estimating causal effects using regression-discontinuity designs (RDDs) in education.

Background

Developed nations depend increasingly on high levels of human capital to compete in sectors of the economy that generate, organize, and use knowledge and data (de la Fuente & Ciccone, 2002; Murnane, Willett, & Levy, 1994). Especially important are skills in math, science, and engineering. The current low standing in mathematics of U.S. students relative to students in other developed nations therefore generates concerns about the long-term prospect for the U.S. economy. Moreover, international studies that compare the entire distribution of achievement reveal substantial inequality in mathematics achievement in the United States as compared with the highest achieving countries (Park, 2013). These studies also reveal alarmingly low performance of the lowest performing U.S. students as compared with the lowest performing students in the highest scoring countries. The lowest performing U.S. students are disproportionately from low-income, minority backgrounds. Thus, increasing achievement among the most disadvantaged young people, who tend to live in large U.S. cities characterized by high levels of poverty and racial segregation (Sampson, 2012), is potentially important to increasing the life chances of those students and to improving the achievement overall in the United States.

Recent Policy Initiatives

Over the past decades, the problem of low math achievement among U.S. high school students has inspired a national movement calling for more rigorous high school course requirements. The National Governors Association Center for Best Practices (2005) has recommended toughening high school graduation requirements, which include 3 to 4 years of mathematics starting with algebra in ninth grade or earlier. A college-prep curriculum for all is now in place in 20 states (Achieve, 2009). The most recently constructed "Common Core State Standards" are also consistent with these goals.

In urban school systems, which enroll the overwhelming majority of low-achieving and disadvantaged U.S. students, recent reform efforts have emphasized engaging all students in academic coursework. For example, in Chicago

where this study takes place, "Algebra for All" in 1997 expanded algebra to low-performing students who traditionally enrolled in remedial arithmetic (Allensworth, Nomi, Montgomery, & Lee, 2009). Subsequently, the "Double-Dose Algebra" policy offered extra time and support to low-achieving students (Nomi & Allensworth, 2009). Programs like Double Dose, in place in nearly half of U.S. urban districts today (Council of the Great City Schools, 2009), respond to the widespread concern that struggling students need extra supports to succeed in academic courses. Other related efforts include the Talent Development Program (Kemple, Herlihy, & Smith, 2005; Mac Iver, Balfanz, & Plank, 1998), offering college-prep courses but with slower pacing for low-achieving students (White, Gamoran, Smithson, & Porter, 1996) and offering transition courses as a bridge to a college-prep coursework (White et al., 1996).

High School Math Reform in Chicago

Prior to 1997, algebra course taking in most Chicago high schools depended largely on a student's incoming math skill: High skill students took algebra and low-skill students took remedial arithmetic. In 1997 and thereafter, all students were required to take algebra. This "Algebra for all" policy not only changed the content of instruction for low-skill students, it also decreased the segregation of classrooms based on prior math skill. Using an interrupted time-series design with nonequivalent control groups, Allensworth et al. (2009) showed that the new "Algebra for All" policy, as intended, dramatically increased algebra enrollment for low-skill students. However, on average, low-skill students' math learning did not improve, plausibly because many students lacked sufficient mathematical background to benefit from instruction in algebra (Allensworth et al., 2009). The policy also led to declines in test scores for high-skill students (Nomi, 2012), suggesting that taking algebra with comparatively low-skill peers had a negative impact for high-skill students.

In 2003, the district launched the "Double-Dose Algebra" policy, a deliberate attempt to overcome the shortcomings of Algebra for All. The Double-Dose policy required all ninth-grade students scoring below the national median on the eighth-grade math test to take two periods of algebra—regular algebra and an algebra support class that focused on building foundational math skills. Also, in an attempt to create consistency between algebra and support coursework, the district recommends that schools offer the two classes sequentially, that the same teacher teach the two classes, and that the same set of students take the two classes together. To follow the district's recommendations, many but not all schools created separate algebra classes—classes composed of Double-Dose students and classes composed of students not assigned to Double Dose. Thus, the Double-Dose policy had just the opposite effect on classroom composition of that generated by the Algebra for All policy. Whereas Algebra for All tended to reduce classroom segregation based on prior peer skill, Double Dose tended to intensify sorting by students' skill levels (Nomi, 2012; Nomi & Allensworth, 2013).

Nomi and Allensworth (2009) used an "RDD" (see Cook, 2008; Thistlethwaite & Campbell, 1960) to assess the impact of the Double-Dose policy on math achievement for students near the cut score by exploiting the cut-score-based course assignment rule using the data on two cohorts of postpolicy students (the 2003 and 2004 cohorts). Their intent-to-treat (ITT) analysis, comparing students scoring just below with those scoring just above the cut point, found a significant positive ITT effect (Nomi & Allensworth, 2009). However, not everyone took the assigned course; thus, the ITT effect underestimated the impact of actually participating in Double-Dose coursework. To estimate the impact of actual program participation, Nomi and Allensworth (2009) used the cut score as an instrumental variable, providing the impact on those who were induced to take the course by virtue of scoring below the cut point, that is, the "complier-average causal effect" (CACE) or sometimes called "local average treatment effect" (Angrist, Imbens, & Rubin, 1996). Earlier studies have found significant positive effects of taking Double Dose on both sort-term and long-term outcomes, including high school graduation and college enrollment (Cortes, Goodman, & Nomi, 2015; Nomi & Allensworth, 2009).

Although these studies revealed a benefit of Double Dose, the interpretation of this finding is ambiguous in light of the fact that the policy

significantly increased classroom segregation (or "sorting") based on prior math skill. Did Double Dose improve students' outcomes strictly because it doubled instructional time? Or did assignment to comparatively low-skill classrooms undermine the impact of policy for students whose skill was near the national median?

Hypotheses and Rationale

Past research gives us reason to predict a negative effect of attending a low-skill classroom. Argys, Rees, and Brewer (1996); Hoffer (1992); and Loveless (2009) found that high-skill students' achievement suffered in detracked schools where classmates' abilities were comparatively low. Nomi (2012) also found a negative effect on high achievers of having lower skill classmates in her evaluation of "Algebra for All" in Chicago. Recent studies on peer effects found positive effects of being assigned to peers with higher skills (Duflo, Dupas, & Kremer, 2011) and negative effects of having low-skill peers (Imberman, Kugler, & Sacerdote, 2012).[1] We caution that the literature does not speak with a single voice on this question. Burris, Hubert, and Levin (2006) studied the impact of expanding the pace of math instruction for all students in mixed-ability middle school classes. Attending a class with lower skill peers did not reduce the learning of higher skill students in that study (see also Boaler & Staples, 2008; Oakes, 2005; B. C. Rubin, 2008).

Researchers have offered several explanations for the observed negative impacts of low classroom peer skill. Most fundamentally, the effect would come from instructional change: Teachers have been found to pitch the level, expectation for content mastery, and pace of instruction to the median level of prior skill of the students in the classroom (Barr & Dreeben, 1983; Pallas, Entwisle, Alexander, & Stluka, 1994). For this reason, higher skill students may tend to be bored in a class with low-skill peers (Rosenbaum, 1999). Student participation and pedagogical practices are also related to academic composition of classrooms; students in high-skill classes are more likely to actively participate in the class and be engaged in discussions (Gamoran, Nystrand, Berends, & LePore, 1995). In contrast, low-track or low-skill classes tend to be more disruptive and have an overall low-quality classroom instructional environment (Oakes, 2005;

Page, 1991; Rosenbaum, 1976; Wheelock, 1992). Classroom management is also problematic, perhaps in part because schools tend to assign less skilled teachers to low-skill classrooms (Kelly, 2004). Finally, assignment to a low-skill class may convey stigma or negative expectations about students' capacity to learn math. If students internalize these expectations, it may tend to undermine their motivation to learn and hence their outcomes (Dweck, 2006; Oakes, 2005; Schafer & Olexa, 1971).

For these reasons, we hypothesize the following for median-skill students:

Hypothesis 1 (H1): The impact of assignment to Double-Dose algebra, as well as taking Double-Dose algebra on those who complied with course assignment, will be small in schools where Double-Dose classrooms are composed of lower skill peers than are Single-Dose classrooms.

Hypothesis 2 (H2): The impact of assignment to Double-Dose algebra, as well as taking Double-Dose algebra on those who complied with course assignment, will be large in schools where being assigned to Double Dose has little or no impact on classroom academic composition.

If H1 and H2 are correct, the impact of the policy may be highly heterogeneous for students in the middle of the prior achievement distribution. Such a result would change our belief about the magnitude of impact that we might achieve with this kind of reform while also raising tricky questions about how best to organize instruction under the policy. We turn to these questions in the concluding section of this article.

Method

Sample and Data

The CPS system 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 use data on 11,296 first-time ninth graders attending 60 nonselective, comprehensive neighborhood public high schools in Chicago during the 2003–2004 academic year (Table 1).[2] The total number of algebra classes is 969 with the average class size of 24 students. We exclude students with disabilities, who comprised 18% of the first-time ninth graders. Many schools exempted students with disabilities from taking Double-Dose algebra and instead assigned these students to take algebra with other students with disabilities. We also excluded students without valid classroom data.[3] 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-price lunch; 55% were African American and 34% were Hispanic (see Table 1).[4]

Our outcome variables are 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, and algebra course grades. 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 scores on the Grade 8 Iowa Test of Basic Skills (ITBS) in math, which is used to determine Double-Dose eligibility. To measure classroom peer achievement, we first created a latent math score, using a vector of ITBS math scores from third through eighth grade.[5] We then computed the average of a student's classmates on this latent math score.

Implementation of Double Dose

Students scoring below the 50th percentile on the ITBS during the spring of their eighth-grade year were expected to take Double-Dose algebra. In fact, 82% of those scoring below the cut point did take Double Dose whereas 4% of those scoring above the cut point and who were therefore not required to take Double Dose did in fact take Double Dose. Thus, compliance with policy, though not perfect, was high as suggested in Figure 1. The horizontal axis of Figure 1 displays the values of student eighth-grade ITBS test scores. The vertical axis is the fraction of

TABLE 1
*Descriptive Statistics (*N *= 11,296)*

Variables	M	SD
Scoring below cut point	.40	.49
Latent math skills	.20	.70
Algebra scores	0	1
Taking Double-Dose algebra	.35	.48
Class peer ability	.18	.58
Free/reduced-price lunch	.86	.35
African American	.55	.50
White	.07	.26
Hispanic	.34	.48
Asian	.03	.18

Note. The analytic sample consists of students without disabilities who attend regular high schools.
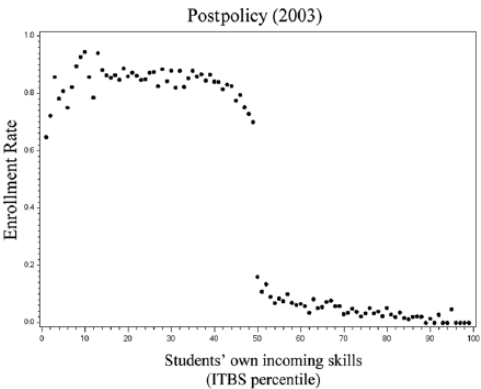

FIGURE 1. *Double-Dose algebra enrollment rate by math percentile scores.*
Note. ITBS = Iowa Test of Basic Skills.

students taking Double Dose. Each dot on the plot is the proportion of students taking Double Dose conditional on their shared ITBS score. Students scoring below the cut point exhibited a very high probability of taking Double Dose whereas those scoring above the cut point were very unlikely to take the course. There is a marked drop in Double-Dose enrollment at the cut point set by the district.[6]

Crucially, scoring below the cut point also substantially reduced, on average, the classroom mean prior skill of one's peers. This dramatic effect on classroom segregation is apparent in Figure 2. The left panel of the figure displays classroom average prior math achievement
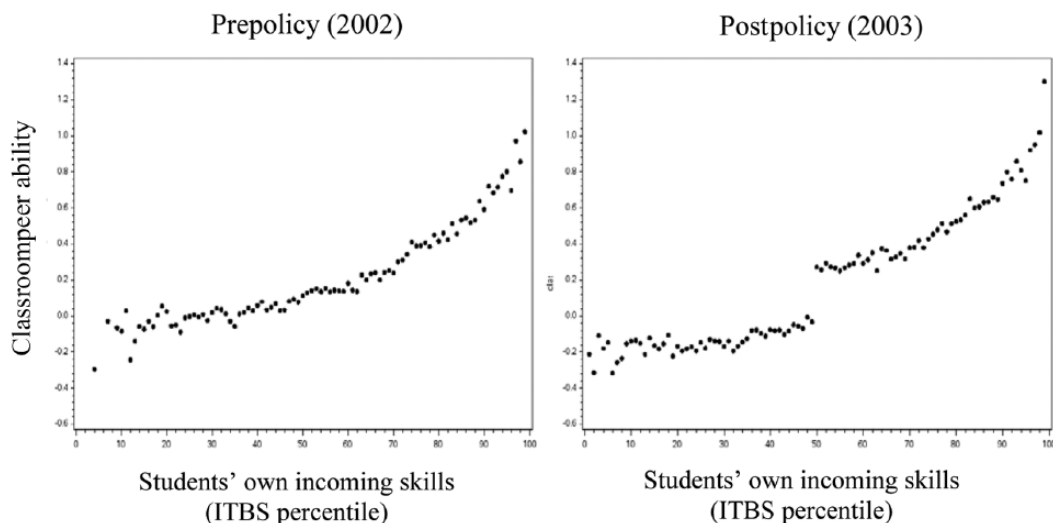
FIGURE 2. *Classroom average skill levels by math percentile scores.*
Note. ITBS = Iowa Test of Basic Skills.

(vertical axis) as a function of a student's own ITBS score (horizontal axis) 1 year prior to the implementation of the Double-Dose policy. The slope of the line describing this association is an index of segregation. To see this, note that a slope near zero would indicate a student's prior skill does not predict the average skill of his or her classmates, that is, classrooms are not segregated based on prior skill. The positive slope of this line indicates that, in the present case, in the year prior to the policy, students with higher incoming math skills tended to have classroom peers with higher math skills.

The right panel of the figure displays the same association, but now during the year of the implementation of the Double-Dose policy. We see a marked discontinuity at the enrollment cut score of 50. On average, scoring below the cut point sharply reduced the classroom mean prior skill of one's peers. This discontinuity can be regarded as the impact of the Double-Dose policy on classroom peer skill, generating, in principle, a powerful natural experiment that enables us to assess the impact of classroom peer skill for students in the neighborhood of the cut point.

In sum, scoring below the cut point on the pretest had two effects, on average: It strongly increased the probability of taking Double Dose and increased the chance of taking algebra with low-skill classmates. If these processes worked the same way in every school,

we would have no basis for separating the impact of increased instructional time from the impact of classroom peer skill. However, policy implementation varied remarkably across schools in both the degree of compliance with the cut-score-based Double-Dose assignment and the degree to which schools created skill-based classroom segregation. This heterogeneity in implementation enables us to separately identify the impact of taking Double Dose from the effect of having low-skill peers under assumptions described in the next section.

Theoretical Model and Analytic Approach

Our primary analysis uses school-specific instrumental variables within a parametric model for the association between the pretest and the outcome to identify, for each school, the discontinuity at the cut point. This approach provides a statistically precise summary of all key causal effects under assumptions that we delineate. We also conduct a nonparametric sensitivity analysis using methods recommended by Imbens and Lemieux (IL; 2008).

The conceptual model for both analyses is displayed in Figure 3. For median-skill students, scoring below the cut point ("T") increases the probability of taking Double-Dose algebra ("D"), on average, by an amount denoted by γ. Taking Double Dose changes the average prior skill of
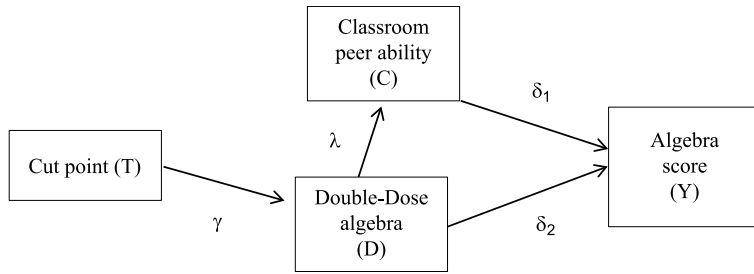
FIGURE 3. *Causal model.*

classroom peers in one's algebra classroom ("C") by λ. The school-specific partial impacts of taking Double Dose and of classroom peer skill on algebra learning are, respectively, δ_1 and δ_2. The ITT effect, denoted by β, is the total effect of scoring below the cut point on algebra achievement. Applying this model to any school j that implements the policy, we therefore can write

$$\beta_j = \gamma_j \delta_{1j} + \gamma_j \lambda_j \delta_{2j} = \gamma_j \delta_{1j} + \theta_j \delta_{2j}. \quad (1)$$

Here $\theta_j = \gamma_j \lambda_j$ is the ITT effect of T on C in school j, that is, the impact of scoring below the cut point on classroom average prior skill. Importantly, this causal process occurs in every school. However, we are interested in the average ITT effects defined as β, γ, θ (the citywide averages of β_j, γ_j, and θ_j, respectively); as well as the overall mean complier average impacts defined as δ_1 and δ_2 (the citywide averages of δ_{1j} and $\delta_{2j} A$, respectively).

Primary Analysis: A School-by-School Parametric Approach

Following an approach devised by Kling, Liebman, and Katz (2007; see also Duncan, Morris, & Rodrigues, 2011), we use 60 school-specific instrumental variables to identify the impact of taking Double-Dose algebra and the impact of classroom peer skill on algebra achievement and grades.

Notation and Model. Let $T_{ij} = 1$ if student i in school j scores below the cut point, and $T_{ij} = 0$ if not; let $D_{ij} = 1$ if that student takes Double-Dose algebra, where $D_{ij} = 0$ if not. Let C_{ij} be the average eighth-grade math test score of the classroom peers of student i in school j. Define X_{ij} as

the pretest score (i.e., ITBS score) for this student; X_{ij} is centered on the 50th percentile, the cut point used in making Double-Dose assignments. We specify the effect of scoring below the cut point on each of our causal variables of interest using a pair of first-stage equations, for each school, as

$$D_{ij} = \gamma_{0j} + \gamma_j \left(T_{ij} - \overline{T}_j \right) +$$
$$\gamma_{1j} \left(X_{ij} - \overline{X}_j \right) + \gamma_{2j} \left(T_{ij} X_{ij} - K_j \right) + e_{dij}, \quad (2a)$$

$$C_{ij} = \theta_{0j} + \theta_j \left(T_{ij} - \overline{T}_j \right) + \theta_{1j} \left(X_{ij} - \overline{X}_j \right) +$$
$$\theta_{2j} \left(T_{ij} X_{ij} - K_j \right) + e_{cij}. \quad (2b)$$

Here $\overline{T}_j$, $\overline{X}_j$, and K_j are the school means, respectively, of the predictors T_{ij}, X_{ij}, and $T_{ij} X_{ij}$.[7] Equations 2a and 2b are piecewise linear. The linear association in school j between X_{ij} and D_{ij} above the cut point is γ_{1j} and below the cut point is $\gamma_{1j} + \gamma_{2j}$. The difference between intercepts defined at the cut point ($X_{ij} = 0$) is the discontinuity γ_j and is, under our assumed piecewise linear specification, equal to the ITT effect, that is, the impact of scoring below the cut point on the probability of taking Double Dose for students whose prior skill is near the cut point. Similarly, the linear association between X_{ij} and C_{ij} above the cut point is θ_{1j} and below the cut point is $\theta_{1j} + \theta_{2j}$. The difference between intercepts defined at the cut point is the discontinuity θ_j, which is the ITT effect on classroom peer skill. Equations 2a and 2b are random coefficient models, also known as hierarchical linear models (Raudenbush & Bryk, 2002). Specifically, in Equation 2a, the coefficients $\gamma_{0j}, \gamma_j, \gamma_{1j}, \gamma_{2j}$ vary randomly around their overall (citywide)

means $\gamma_0, \gamma, \gamma_1, \gamma_2$ with variances and covariances to be estimated, and similarly for the coefficients $\theta_{0j}, \theta_j, \theta_{1j}, \theta_{2j}$ in Equation 2b which vary randomly around their overall (citywide) means $\theta_0, \theta, \theta_1, \theta_2$.

Our theoretical model for the impact of Double Dose and classroom peer skill on algebra achievement has a similar form

$$
\begin{aligned}
Y_{ij} = \phi_{0j} &+ \delta_{1j}\left(D_{ij} - \bar{D}_j\right) + \\
&\delta_{2j}\left(C_{ij} - \bar{C}_j\right) + \phi_{1j}\left(X_{ij} - \bar{X}_j\right) + \\
&\phi_{2j}\left(T_{ij}X_{ij} - K_j\right) + e_{yij},
\end{aligned} \tag{3}
$$

except that we cannot now estimate school-specific values δ_{1j} and δ_{2j}. However, we can estimate their average values δ_1 and δ_2.[8] In Equation 3, the coefficients $\phi_{0j}, \phi_{1j}, \phi_{2j}$ vary randomly around their overall means ϕ_0, ϕ_1, ϕ_2 with variances and covariances to be estimated. Direct estimation of Equation 3 would produce biased estimates of the causal effects δ_1 and δ_2 of Double Dose (*D*) and classroom peer ability (*C*), respectively, if unobserved background variables that predict the outcome also influence Double-Dose participation and/or classroom peer skill conditional on the pretest. However, under our RDD and the assumptions described below, we can predict $(D_{ij} - \bar{D}_j)$ and $(C_{ij} - \bar{C}_j)$ without bias by estimating Regression Equations 2a and 2b. We can then substitute these predicted values into Equation 3 to obtain unbiased estimates of δ_1 and δ_2 as is standard in two-stage least squares.

Reduced Form. To see how the procedure just described corresponds to the parameters described in Figure 3 (and Equation 1), we can derive the "reduced form," that is, the model for ITT effect of *T* on *Y*, often called the "total effect" of *T* on *Y*. We obtain the reduced form by substituting expressions for $(D_{ij} - \bar{D}_j)$ and $(C_{ij} - \bar{C}_j)$ into our theoretical model (Equation 3) and taking an expectation, with the result that

$$
\begin{aligned}
E\left(Y_{ij}\right) = \beta_0 &+ \beta\left(T_{ij} - \bar{T}_j\right) + \\
&\beta_1\left(X_{ij} - \bar{X}_j\right) + \beta_2\left(T_{ij}X_{ij} - K_j\right),
\end{aligned} \tag{4}
$$

where $\beta = \gamma\delta_1 + \theta\delta_2$ is the average ITT effect on *Y*. Note also that $\beta_0 = \phi_0 + \delta_1\gamma_0 + \delta_2\theta_0$; $\beta_1 = \phi_1 + \delta_1\gamma_1 + \delta_2\theta_1$; and $\beta_2 = \phi_2 + \delta_1\gamma_2 + \delta_2\theta_2$.

Assumptions

Reardon and Raudenbush (2013) derived the assumptions that must be met in multisite randomized trials that use site-specific instruments to identify the impact of multiple mediators of a treatment. We describe how these apply in our RDD study in Appendix A (available in the online version of the journal) in detail and provide a brief summary here.

Stable Unit-Treatment Value Assumption (SUTVA). Under this assumption (D. B. Rubin, 1986), each participant possesses a single potential outcome under each possible treatment assignment. This implies that a participant's potential outcome does not depend on the treatment assignment of one's peers (D. B. Rubin, 1986). This assumption would appear naïve in an analysis that considers only the impact of Double Dose because the impact of scoring below versus above the cut point may well depend not only on taking Double Dose but also on the treatment assignment of other students. Thus, following Hong and Raudenbush (2006), we have effectively modified the conventional SUTVA to include the possible influence of classroom mean skill on potential outcomes.

Exclusion Restrictions. We assume that scoring below the cut point cannot influence classroom peer skill except by inducing a student to take Double Dose (i.e., no direct path between *T* and *C* in Figure 3); scoring below the cut point cannot influence algebra learning except by changing coursework or classroom peer skill (no direct path between *T* and *Y* in Figure 3). At one level, these exclusion restrictions are quite reasonable: There is little reason to think that scoring below versus above the cut point would have much influence unless it affects the setting in which the child learns or the course content. However, the interpretation of the effect of "classroom peer skill" is open to considerable debate. A negative effect might reflect not only peer influences but also institutional processes that influence the

cognitive level, amount, or quality of instruction occurring in classrooms that vary in prior skill as described in our section on hypotheses above. We revisit this in the concluding section.

No Confounding of Treatment Assignment. This assumption is satisfied in the RDD study because treatment assignment is known (i.e., determined by the observed pretest scores). However, we must assume that the functional form of association between the variable that determines treatment assignment and the outcome is correctly specified. Although our graphical analyses shown in the "Findings" section suggest that the piecewise linear model (Equations 2–4) adequately captures the pretest and outcome relationship, we relaxed this assumption in our nonparametric sensitivity analysis.

Linearity and Additivity of Classroom Peer Effects. Our theoretical model is a linear additive function of Double-Dose participation and classroom peer effects (Equation 1). We checked these assumptions graphically and in our sensitivity analyses.

Monotonicity. We assume that scoring below the cut point cannot reduce the probability of taking Double Dose or increase classroom peer skill. We regard these as plausible assumptions, and ones that accord with our school-by-school graphical analyses.

The next two assumptions are required for estimating the independent impact of Double Dose and classroom peer skill.

The Association Between Our Instrument and at Least One of Our Two Endogenous Variables (Double Dose and Classroom Peer Skill) Must Vary Across Schools. This is demonstrated in the result section, ITT Impact on Taking Double-Dose Algebra and Classroom Peer Skill.

The School-Specific Impact of Scoring Below the Cut Point on Double-Dose Enrollment Cannot Be Perfectly Correlated With the School-Specific Impact of Scoring Below the Cut Point on Classroom Peer Composition. As discussed in the result section, ITT Impact on Taking Double-Dose Algebra and Classroom Peer Skill, this

assumption is met with the overall correlation of $r = -.33$.

Independence of Site-Mean Compliance and Effect Assumptions. Identification of our theoretical model (Equation 1) using Equations 2 to 4 depends on several additional assumptions regarding the relationship between school-specific implementation of the policy and the impact of implementation. With regard to course compliance, we must assume that schools that comply Double-Dose policy (and therefore have large values of γ_j) do not have students who stand to benefit more (or less) from complying (as indicated by the impacts δ_{1j} and δ_{2j}). For example, we might consider the following possibility. Suppose that, in one school, those scoring below the cut point would gain a lot from Double Dose whereas those scoring above the cut point would be better off with Single-Dose algebra. If school personnel knew that, course compliance might be high in such a school. If, in another school, students on both sides of the cut point would benefit (or would be hurt) by taking Double Dose, school personnel would be rational by failing to comply. If this kind of rationality is held across all schools, compliance would be associated with the impact of complying. Similarly, for classroom peer sorting, we must assume that the site-mean impact of scoring below the cut point on classroom peer skill is not associated with δ_{1j} and/or δ_{2j}. These scenarios do not seem entirely plausible, but these assumptions are not directly testable. Instead, we use our sensitivity analyses to check on the possibility that failure of this assumption has distorted our results.

Findings

ITT Impacts

ITT Impact on Outcomes.[9] Controlling for the pretest (see Equation 4), scoring below the cut point increased algebra achievement, on average, by an estimated standardized effect size of $\hat{\beta} = .067$ ($SE = .027$). This is the average ITT effect, slightly smaller than that estimate by Nomi and Allensworth (2009) who used a global quadratic specification rather than our local linear specification. We also see some evidence that these ITT

impacts are heterogeneous across the 60 neighborhood high schools: We estimate the standard deviation of the ITT effect to be .07, although this estimate is not precise when other predictors in the model are allowed to have random coefficients. Nonetheless, this result suggests that the 95% plausible value interval (Raudenbush & Bryk, 2002) for school-specific ITT effects ranges from −.07 to .20 *SD*.

ITT Impact on Taking Double-Dose Algebra and Classroom Peer Skill. To what extent does scoring below the cutoff point induce shifts in the probability of taking Double Dose and classroom peer skill? To answer this question, we estimated Models 2a and 2b. Both effects are strong, on average, but vary substantially and significantly from school to school (see Table 2). Specifically, we estimate the average effect of scoring below the cut point on Double-Dose enrollment to be $\hat{\gamma} = .72$ (*SE* = .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, on average, by about 72%, clearly a large effect. However, this effect also varies substantially from school to school, $\hat{\tau}_\gamma^2 = .046$, $\chi^2(59) = 552.73$, $p = .00$, a fact that is illustrated graphically in Figure 4, which displays school-by-school empirical Bayes estimates on the horizontal axis.

For students whose true prior achievement is around the cut point, scoring below the cut point also induces a substantial reduction in classroom peer skill, on average. Across all 60 schools, the average impact is estimated to be $\hat{\theta} = -.27$ (*SE* = .03) standard deviations on the scale of prior math skills (see column 3, Table 2). Moreover, this effect varies significantly and substantially from school to school, $\hat{\tau}_\theta^2 = .04$, $\chi^2(59) = 384.64, p = .00$. This variation is illustrated graphically by the vertical axis in Figure 4, displaying school-by-school empirical Bayes estimates of θ_j. The large variance of this effect implies that in some schools, scoring below the cut point has little or no effect on classroom peer skill, whereas in other schools, the effect is very large. To see this, let us compare schools that are predicted to produce an effect of 1 *SD* above average to schools that are predicted to produce an effect of 1 *SD* below average. These predictions

are $\hat{\theta}_j = \hat{\theta} \pm \sqrt{.04} = (-.07, -.47)$, indicating only a small effect of −.07 for one group and a very large effect of −.47 for the other. The difference of −.40 between these two effects is a standardized effect size in terms of individual students' peer skill levels as we first standardized the latent math skill variable (*SD* = 1) to construct average classroom peer ability.[10]

As Figure 4 shows, these two effects are not strongly correlated with an estimated correlation coefficient of $\text{Corr}(\gamma_j, \theta_j) \approx -.33$. Thus, the large variation across schools underscores our conclusion that we have 60 independent natural experiments in which scoring below the cut point induces varying degrees of change in classroom peer composition.

Instrumental Variable Results

Effects of Double Dose and Classroom Peer Skill. Our analysis using the instrumental variable method (Equations 2 and 3) finds a significant positive impact of taking Double-Dose algebra, $\hat{\delta}_1 = .196$ (*SE* = .05, *t* = 5.24, *p* < .001; Table 3). This implies that, if scoring below the cut point had no impact on peer academic composition, taking Double-Dose algebra would improve algebra scores by about .20 *SD*s. Holding constant Double Dose, classroom peer skill significantly predicts increased achievement, $\hat{\delta}_2 = .221$ (SE = .89, *t* = 3.04, *p* < .01), that is, an increase in classroom average ability by 1 *SD* is associated with an improvement in algebra scores by .22 *SD*.

These results suggest that the ITT impacts of the policy depend on both the level of course compliance and the degree of sorting, each of which varied considerably across schools. To graphically present the ITT effects for schools with different levels of course compliance and degree of sorting, we plotted the expected ITT effects along with their 95% confidence intervals (the *y*-axis) by the degree of sorting (θ) for schools with moderate course compliance rate of .6 (Figure 5) and high course compliance rate of .9 (Figure 6). These figures suggest that, given the level of course compliance, the ITT effects on algebra scores are larger for schools that did not segregate classrooms by students' pretest scores. Also, comparing moderate-compliance schools (Figure 5) and high-compliance schools (Figure 6),

TABLE 2

Effect of Cut Point on Double-Dose Algebra Enrollment and Classroom Academic Composition: All 60 Schools (N = 10,131)

	Algebra scores		Double-Dose algebra enrollment		Classroom academic composition	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Intercept	−.209***	.035	.436***	.024	.026	.035
Below cut	.067*	.027	.718***	.030	−.274***	.029
ITBS	.025***	.001	−.002***	.001	.012***	.001
Below Cut × ITBS	−.010***	.002	.001	.002	−.005***	.001
	Variance component	$\chi^2\ df = 59$	Variance component	$\chi^2\ df = 59$	Variance component	$\chi^2\ df = 59$
Intercept	.070***	1,426.000	.035***	6,349.37	.072***	7,969.809
Below cut	.005	57.099	.046***	552.73	.041***	384.649
ITBS	.000**	90.224	.000***	226.07	.000***	439.570
Level 1 σ^2	.497		.073		.085	

Note. Students whose ITBS scores are above the 75 percentile are excluded from the analysis. ITBS = Iowa Test of Basic Skills.
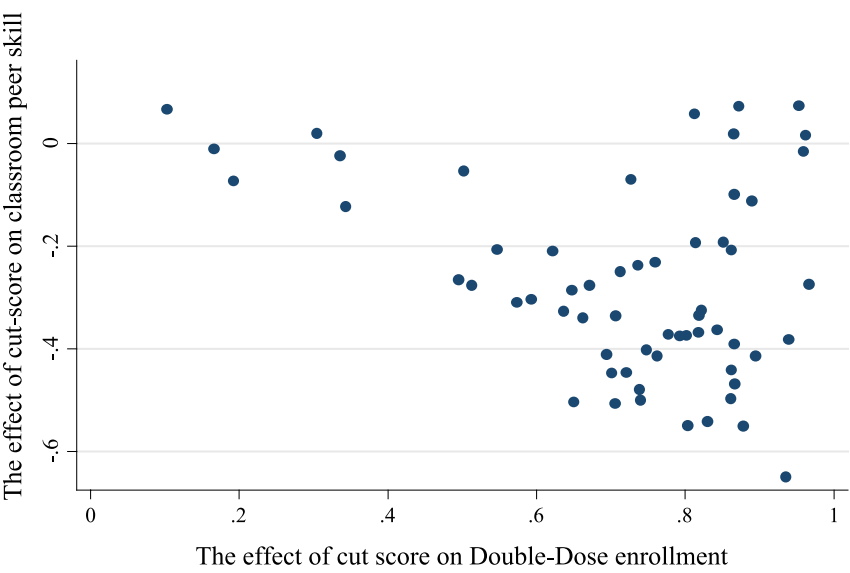*$p < .1$. **$p < .01$. ***$p < .001$.

FIGURE 4. *The effect of cut score on Double-Dose algebra enrollment by the effect of cut score on classroom peer ability.*

we see that the ITT effects are larger for high-compliance schools with the same level of sorting (θ).

CACEs. Our results imply that, for compliers, the impact of taking Double Dose on classroom peer skill was $\hat{\lambda} = -.392$ on average. The predicted impact of taking Double Dose on the outcome for these students was $\hat{\delta} = \hat{\delta}_1 + \hat{\lambda}\hat{\delta}_1 = .196 - .382 \times .221 = .112$ on average. For compliers attending a school in which scoring below the cut point reduced classroom

441

Estimated Effect of Classroom Peer Ability, Controlling for Double-Dose Algebra and Effect of Double-Dose Algebra, and Controlling for Classroom Peer Ability on Algebra Scores

	Algebra scores			
	All schools		High-compliance schools	
	Coefficient	SE	Coefficient	SE
Intercept	−.209***	.035	−.227***	.043
Double Dose	.196***	.047	.166***	.042
Classroom peer ability	.221*	.090	.166*	.080
ITBS	.023***	.002	.024***	.002
Below Cut × ITBS	−.010***	.002	−.010***	.002
	Variance component	$\chi^2\ df = 59$	Variance component	$\chi^2\ df = 59$
Intercept	.071***	1,425.287	.073***	1,060.849
ITBS	.000**	98.467	.000*	61.960
Level 1 σ^2	.497		.501	
Total N	10,131		6,913	

Note. Students whose ITBS scores are above the 75 percentile are excluded from the analysis. ITBS = Iowa Test of Basic Skills.
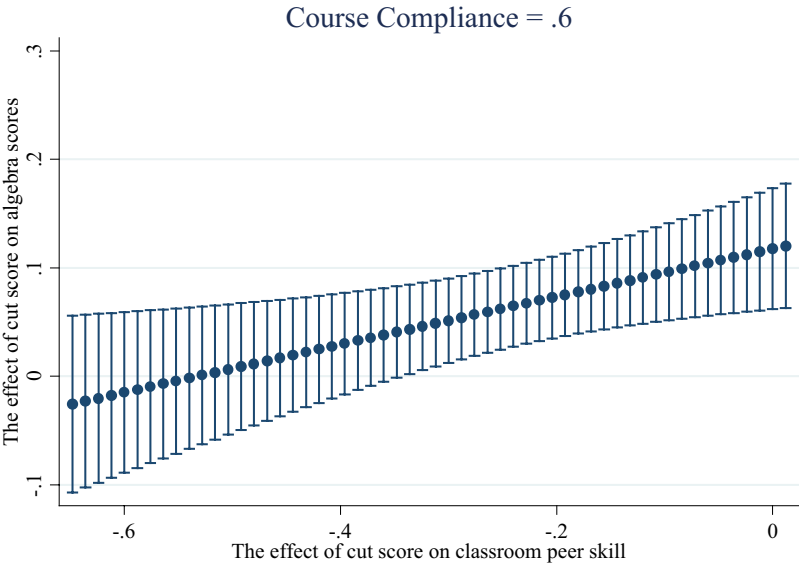*$p < .1$. **$p < .01$. ***$p < .001$.

Course Compliance = .6

FIGURE 5. *Predicted ITT effects on algebra scores by the effect of cut score on classroom mean peer skill for moderate-compliance schools (course compliance rates of .6).*
Note. The value of the effect of cut score on classroom peer skill is based on the 95% plausible interval. ITT = intent to treat.

peer skill by .47 (1 *SD* greater reduction than average), our model predicts $\hat{\lambda} = -.655$ so the impact of taking Double Dose on the outcome is

$\hat{\delta} = \hat{\delta}_1 + \hat{\lambda}\hat{\delta}_1 = .196 - .665 \times .221 = .049$. For compliers attending a school in which scoring below the cut point reduced classroom peer skill by
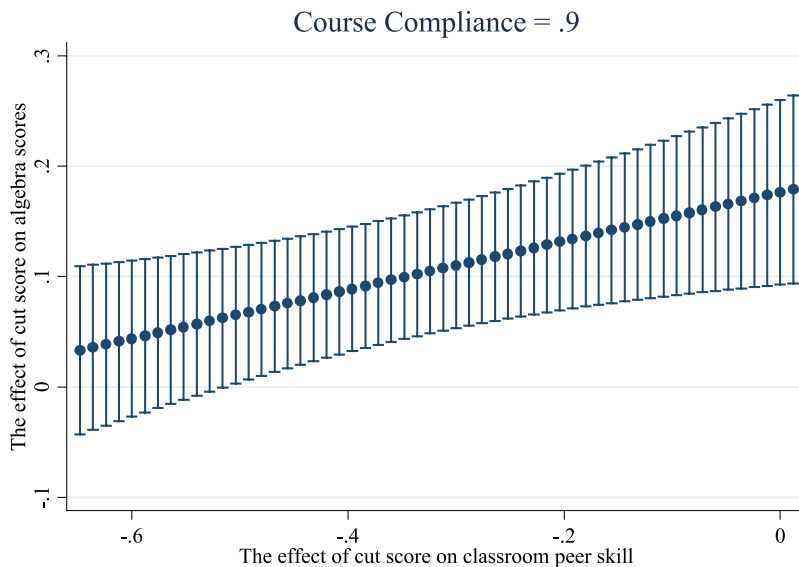
FIGURE 6. *Predicted ITT effects on algebra scores by the effect of cut score on classroom mean peer skill for high-compliance schools (course compliance rates of .9).*
Note. The value of the effect of cut score on classroom peer skill is based on the 95% plausible interval. ITT = intent to treat.

only .07 (1 *SD* smaller reduction than average), our model predicts $\hat{\lambda} = \hat{\theta}/\hat{\gamma} = -.07/.718 = -.097$ so the expected impact of taking Double Dose is $\hat{\delta} = \hat{\delta}_1 + \hat{\lambda}\hat{\delta}_1 = .196 - .097 \times .221 = .175$. These results are corroborated by our nonparametric analysis (see below). Figure 7 illustrates the relationship between the impact of taking Double Dose on classroom peer skill (horizontal axis) and the CACE (vertical axis).

Results for Course Grades

We conducted the same analysis using algebra course grades as an outcome. Prior research by Nomi and Allensworth (2009) showed that Double-Dose algebra led to higher algebra grades. Consistent with their findings, we found a statistically significant ITT effect; on average, scoring below the cut point would lead to higher algebra grades by .26 (*SE* = .05, *t* = 5.09, *p* < .001). Also, controlling for classroom peer ability, the average effect of taking Double-Dose algebra is .19 although this did not reach statistical significance (*SE* = .12, *t* = 1.60, *p* = .110). However, controlling for Double-Dose enrollment, we found a negative effect of classroom peer ability on algebra course grades, $\hat{\delta}_2 = -.37$ (*SE* = .20, *t* = −1.809, *p* < .1), suggesting that

students tend to have higher grades in classes with lower skill peers.

Sensitivity Analyses

We conducted sensitivity analyses to see whether our results were sensitive to failure of the "independence of site-mean compliance and effect" assumption, the assumption of a linear additive association between *C* and *Y*, and the functional form assumption of our parametric model. Here, we present only the result on the impact of classroom peer skills—the key causal impact of this study that relies on all of these assumptions. The complete sensitivity analyses are presented in Appendix B (available in the online version of the journal). Overall, the results strongly corroborated the findings just described.

Sensitivity of Inferences About the Impact of Classroom Peer Skill on Algebra Achievement. Our parametric analysis isolated the impact of classroom peer skill by parametrically identifying ITT effects on Double-Dose enrollment and classroom peer skill school by school (Equations 2a and 2b). To check the sensitivity of our estimate to the violation of three assumptions
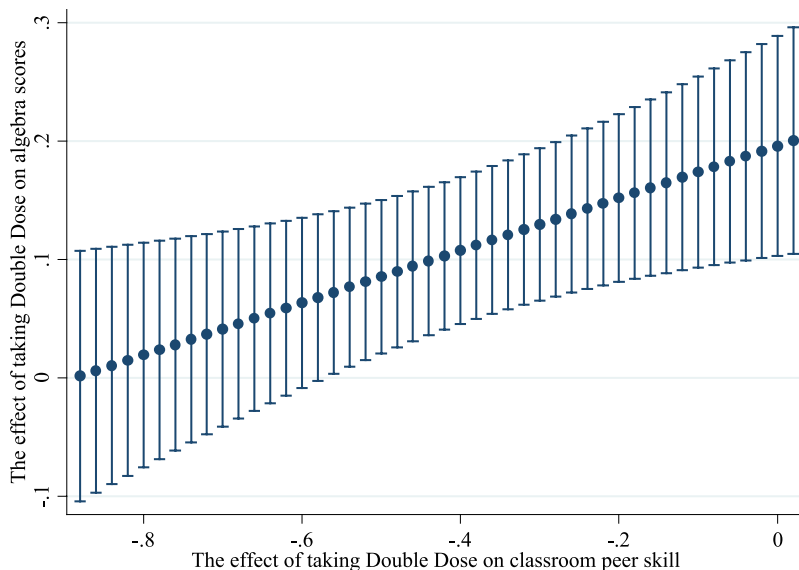
443

FIGURE 7. *Predicted CACE on algebra scores by the effect of taking Double Dose on classroom mean peer skill for high-compliance schools.*
CACE = complier-average causal effect.

TABLE 4
Number of Schools by Compliance Status

| | Course compliance | | | |
Sorting	Low ($\gamma^*_j < .4$)	Middle ($.4 < \gamma^*_j < .7$)	High ($\gamma^*_j > .7$)	Total school N
Low ($\theta^*_j > -.25$)	6	10	15	22
High ($\theta^*_j < -.25$)	0	3	26	38
Total school N	6	13	41	60

stated above, we estimate the ITT effect and CACE—the effect of scoring below the cut point and taking Double-Dose algebra on the outcome—within a given level of course compliance γ by using nonparametric RDD as recommended by IL (2008). We then compare these impacts among schools that are similar in course compliance γ, but differ in the degree of skill-based sorting θ (i.e., high vs. low sorting). Note that these ITT and CACE estimates do not rely on the linear and additivity assumption of Equation 3 or the mean independence assumption.[11] Nor does the IL method depend on the parametric RDD model of Equations 2 and 3. Thus, if all other RDD assumptions hold, the difference in the ITT and CACE estimates between two groups of schools with the same level of

course compliance, but differing in the degree of skill-based sorting, is attributable to the difference in the degree of sorting.

However, the IL method uses only the data on students who scored near the cut point. As a result, we do not have enough data to estimate ITT effects school by school. Instead, we stratified schools based on parametric estimates of γ and θ (Table 4). Pooling the data within strata, we estimated these same effects using IL and found very similar estimates of γ and θ (see Appendix B, available in the online version of the journal). Next, we estimated stratum-specific ITT and CACE on algebra scores within the strata of the 41 high-compliance schools ($\hat{\gamma} \geq .70$). It is among these high-compliance schools that the degree of sorting varies the most as is depicted

TABLE 5

Comparisons of Parametric and Nonparametric Estimates of ITT and CACE Among High-Compliance Schools

	Parametric		Nonparametric	
	High sorting	Low sorting	High sorting	Low sorting
ITT	.061	.111	−.001	.146
	NA	NA	(.070)	(.100)
CACE	.073	.132	−.001	.191
	NA	NA	(.095)	(.152)

Note. Parametric estimates are the pooled average impacts across schools within stratum (i.e., high-sorting and low-sorting schools with high course compliance). Standard errors are in parenthesis. ITT = intent to treat; CACE = complier-average causal effect.

in Figure 4. Among them, 26 schools were classified as "high-sorting" schools ($\hat{\theta} < -.25$) and 15 schools as "low-sorting" schools ($\hat{\theta} > -.25$); the average $\hat{\theta}$ was −.439 for high-sorting schools and .095 for low-sorting schools. In contrast, the average Double-Dose compliance rates were nearly identical with the average $\hat{\gamma}$ of .841 for high-sorting schools and .832 for low-sorting schools.

The estimated ITT effects on algebra scores within the 10 percentile bandwidths using the IL method were .146 for low-sorting schools and −.001 for high-sorting schools; estimated CACEs were .192 and −.001 for low- and high-sorting schools (Table 5).[12] These point estimates are qualitatively similar to those estimated by our parametric model but are less precise.

Discussion

Our study can be regarded as a synthesis of 60 independent natural experiments. The Double-Dose policy mandated that all students take academic algebra in ninth grade, but that students scoring below the national average on the eighth-grade math test take two periods of ninth-grade math: one period of academic algebra and a second period of math coursework designed to support algebra learning of students who had fallen behind. We found remarkable heterogeneity in how these 60 schools implemented the policy. In most but not all schools,

compliance with assignment to Double-Dose algebra was reasonably high. However, the schools varied enormously in the extent to which implementation induced classroom peer segregation based on prior math skill. This heterogeneity enabled us to disentangle the effects of access to course content via increased instructional time from the effects of classroom peer math skill.

Our evidence shows that for students with median skill in Chicago neighborhood high schools, the overall impact of the Double-Dose policy depended on the expansion of instructional time afforded by Double-Dose algebra and on the tendency of the policy to assign median-skill students to low-skill classrooms. For those median students, the impact of Double Dose on algebra learning had their classroom peer skill unchanged can be quite substantial. Specifically, those induced to take Double Dose by virtue of scoring below the cut point gained about .20 *SD*s in math achievement in schools where the policy did not segregate low-skill students in Double-Dose classrooms. In contrast, the benefit of Double Dose was very small or null when taking Double Dose meant attending a class with low-skill peers. The effect size of .20 in algebra is substantial, given that the average annual gain from Grades 9 to 10 is reported to be .25 *SD* on the nationally normed tests measuring broader content knowledge in mathematics (Hill, Bloom, Black, & Lipsey, 2007).[13]

It appears that careful attention to course scheduling can enhance the effectiveness of expanding instructional time for median-skill students. This reasoning is based on further analyses that explored whether "mechanical factors" strongly determine the degree to which schools created sorted algebra classes. Specifically, we used the following school-level variables to predict the school-specific impact of scoring below the cut point on classroom peer ability: the school's prior test score mean, prior test score standard deviation, neighborhood disadvantage,[14] racial–ethnic composition, fraction of children in special education, and compliance on Double Dose. Our descriptive analysis showed that, on average, schools that created greater sorting tend to have somewhat lower school mean pretest scores, to enroll a larger percentage of African American students and a smaller

percentage of Hispanic students, and to have a higher concentration of poverty. However, these predictors together explained less than 17% of the variance in the ITT impact on classroom peer ability.[15] The fact that such mechanical factors do not determine the degree of sorting associated with implementation of the policy implies that there is a role for school administrators in designing the Double-Dose schedule, and hence shaping the impact of the policy.

A limitation of our design is that our findings apply only to those near the median of the pretest distribution. However, we find that this group is likely to be a fairly large segment of the high school population of interest here. Using data on prior achievement, ethnicity, and social background, we estimate that roughly one quarter of the students had near zero probability of scoring above the cut point whereas a similar number had virtually no chance of scoring below the cut point. Our findings do not provide information on how these very low-scoring students (who have negligible chance of scoring above the cut point) or very 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. Yet, in our data, nearly 50% of all students have nonnegligible chance of scoring either above or below the cut point. This reflects the fact that the pretest score measured at one time point (i.e., eighth grade) contains measurement error.

Although providing insight into the varied impacts of the Double-Dose policy, our results also contribute to the literature on the impact of classroom peer skill on math learning in secondary school. We found a quite substantial negative impact on median-skill students of taking algebra with low-skill peers. This finding is consistent with Nomi's (2012) finding that high-skill students lost ground when assigned to low-skill classrooms under the Algebra for All policy. Past research suggests that a negative impact of attending a low-skill class might result from several plausible mechanisms, including not only peer effects but also institutional effects.

One plausible explanation for such effects is that the average prior skill of classmates constrains the pace and conceptual level of math instruction that a teacher can provide, at least under conventional pedagogical approaches (e.g., Barr & Dreeben, 1983; Gamoran, 2010; Oakes, 2005; Page, 1991). This might be particularly true if the teacher engages in whole-class instruction, the norm in high school math instruction. In addition, Gamoran et al. (1995) found that students in classes with higher achieving students are more likely to engage in discussions and student-initiated activities. There is also evidence that school leaders often assign comparatively low-skill teachers to low-skill classrooms (Kelly, 2004) or that low-skill peers become discouraged and display low motivation for learning that negatively influences the classroom climate (e.g., Oakes, 2005; Page, 1991). There is also good reason to suppose that assignment to a low-skill class can convey negative expectations for math learning to those so assigned. If such students internalize a negative stigma, their motivation learning may decline, leading to lower math achievement (Oakes, 2005; Schafer & Olexa, 1971).

However, our results showed that the effect of low-skill classroom peers is opposite for algebra course grades, that is, for students with the median skills, their grades tend to be higher in classes with low-achieving peers. This may be because these median-skill students would be the highest achieving students in the Double-Dose algebra class if schools segregated students on the basis of the cut point (i.e., "Fish-pond effects") and/or these students exceed expectations in low-skill classrooms if teachers set different expectations according to overall students' skill levels (e.g., Kelly, 2008).

A key question with regard to algebra content mastery (measured by test scores) that we have not answered is whether very low-skill students benefit from taking Algebra with higher skill peers. If they do, school leaders face a difficult trade-off unless new methods of instruction are found that can optimize learning for students of heterogeneous skill. Prior research has demonstrated some evidence of successful detracking (e.g., Burris et al., 2006); however, this evidence typically comes from high-achieving and well-resourced schools as compared with a typical school in an urban district like Chicago. Thus, generalization of this research, particularly in understanding how very low-achieving students

would respond, may be quite limited.[16] However, if low-skill students do as well in homogeneous classes as in heterogeneous classes with higher achieving peers, then our findings, combined with those of Nomi (2012), would seem to suggest that math instruction works best overall when con-duced in classrooms composed of students who are homogeneous in prior skill. Unfortunately, the RDD method cannot assist us in studying the impact of classroom peer skill on math learning for students with very low or high skills.

Our results also have implications for RDD as a strategy for obtaining valid causal inference. When a pretest determines assignment to a treat-ment, we have an unusually strong quasi-experi-ment. For those induced to take up a treatment by scoring below a cut point, we can rule out the stan-dard concern that unobserved covariates are gen-erating selection bias. However, when the treatment is administered in a group setting (e.g., schools), we can anticipate that the cut-score-based treatment assignment is likely to cluster low-scoring students together for instruction. This could, in turn, influence the nature of the interven-tion. Gibbs (2015) brought this problem to light in her study of half- versus full-day kindergarten. Her multidistrict randomized trial showed a posi-tive effect of full-day kindergarten on emerging literacy skills, with particularly pronounced effects for disadvantaged children. In contrast, in districts using the cut-score-based assignment approach, where only low-income students were provided full-day kindergarten, the impacts were nonsignificant. A plausible explanation is that, in the "means-tested" RDD study, low-income stu-dents were clustered together for instruction, which was not the case in the randomized study where the intervention applied universally, regard-less of income. In contrast, the randomized trial was a universal intervention in which such cluster-ing of low-income children did not occur.

Technically, an RDD study like ours and that of Gibbs is two dimensional: Falling below a cut point not only increases instructional time, but it also shapes the peer composition of the social setting in which instruction occurs. In these stud-ies, it becomes important to disentangle different sources of influence.

The broad theoretical lesson in our study seems important: Instructional time and class composition jointly shape learning opportunities and therefore influence the distribution of human capital during adolescence.

Notes

1. There are numerous literatures on "peer effects" in economics of education, and findings are not easily summarized (e.g., Abdulkadiroğlu, Angrist, & Pathak, 2014; Ammermueller & Pischke, 2009; Angrist & Lang, 2004; Duflo, Dupas, & Kremer, 2011; Hanushek, Kain, Markman, & Rivkin, 2003; Hoxby & Weingarth, 2005; Imberman, Kugler, & Sacerdote, 2012; Lavy, Silva, & Weinhardt, 2012). Many studies are based on elementary schools, and findings generally show posi-tive effects of having high-achieving peers.

2. This study does not use the 2004–2005 cohort because the policy implementation changed in the sec-ond year of the policy. For example, the district no lon-ger recommended algebra and support coursework be offered sequentially during the day, and many schools decided to offer two separate courses. Also course compliance and the degree of sorting differed for the 2004 cohort than for the 2003 cohort. We focused our analysis in the first year of the policy for simplicity.

3. Less than 1% of students are excluded due to this restriction.

4. 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 compile administrative records on all students for each semester since 1991.

5. Students' Iowa Test of Basic Skills (ITBS) scores were first equated through Rasch analysis to remove form and level effects. Then, a two-level hier-archical linear model, nesting years within students, modeled each student's learning trajectory; Level one 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.

6. We do not know the precise reasons for noncompliance; however, the staff at the district office noted that this may be because of scheduling difficulties and decisions made by school staff or students in terms of who need to, or need not to, enroll in Double-Dose courses. Also, preliminary analyses examined how school-level compliance is related to school characteristics, including school's prior test score mean, prior test score standard deviation, socioeconomic status (SES) characteristics, racial–ethnic composition, and fraction of students with identified disabilities. These variables did not predict the level of course compliance.

7. Raudenbush (2009) showed that a random effects model in which predictors are school mean centered gives estimates of the regression coefficients that are identical to those obtained from a school-fixed effects model. However, this approach, called "adaptive centering" with random effects, is flexible in allowing coefficients to randomly vary over schools. This is important in our case to avoid the strong assumption that the relationship between the pretest and the outcome is the same in every school. For this reason, the error terms e_{dij} and e_{cij} include school-specific random coefficients associated with predictors $T_{ij} - \overline{T}_j$, $X_{ij} - \overline{X}_j$ and $T_{ij}X_{ij} - K_j$.

8. Each school has two causal effects δ_{1j} and δ_{2j} but only one instrument T_{ij}. Hence, we cannot identify the separate impacts of Double Dose and classroom peer skill for each school. However, looking across the entire sample of 60 schools, we have 60 instruments T_{ij}, enabling us to estimate two *average* impacts δ_1 and δ_2 under the assumptions we delineate.

9. All of our parametric analyses exclude students who scored above the 75th percentile (very low-achieving students were already excluded from the analysis as most of them are students with disabilities). We made this decision because, as shown in the "Findings" section, the piecewise linear model fits the data very well by imposing this restriction. Also, keeping as much data is critical for estimating school-specific impacts with precision. In contrast, our preliminary analyses showed that the models with higher order polynomials, or higher order polynomials with interactions with the cut point, using the entire sample showed poorer fit to the data.

10. It is also noted that the average skill level, measured by latent math scores, of students just below the cut point is $-.035$ *SD*. Our model suggests that the average class mean peer skill for students just below the cut point is $-.35$ *SD*, considerably lower than their own skill levels. The 95% plausible value interval of class mean peer ability for students just below the cut point ranges from $-.74$ to $.04$.

11. These estimates do not rely on the independence of site-mean compliance and effect assumptions regarding course compliance γ as we are comparing the intent-to-treat (ITT) and complier-average causal effect (CACE) estimates among schools with the same level of γ, on average.

12. For students who scored just below the cut point, predicted classroom mean peer ability is much lower in high-segregation schools ($-.463$ *SD*) than in low-segregation schools ($-.097$ *SD*). This suggests that scoring below the cut point will lead to exposure to peers with much lower skills in high-segregation schools than in low-segregation schools. This is consistent with the hypothesis that, given course compliance, exposure to low-skill peers tends to diminish achievement. It is also noted that observed school characteristics (school mean ability, racial composition, sociodemographic characteristics, and percentage of students with disabilities) are not strongly related to the degree of segregation or course compliance levels (also see the "Discussion" section).

13. In addition, potential costs of taking additional supplemental algebra are likely to be minimal. For example, the total number of core academic courses was not affected by the policy as supplemental algebra was likely to replace an elective course (Cortes, Goodman, & Nomi, 2015). Also, our additional analysis showed that scoring below the cut score had no discernable effect on attendance rates in algebra class or overall attendance rates.

14. A student's concentrated neighborhood disadvantage is a factor composed of four census-based measures of the neighborhood block group in which that student resides: percent unemployed, percent poor, percent receiving welfare, and percent single-parent families. School mean disadvantage is the mean of these measures taken over all students in the school.

15. We also compared classroom mean ability of students just above and below the cut point between high- and low-achieving schools among 41 high-compliance schools. For the 20 high-achieving schools, classroom mean ability for students just above the cut point was .14 and that for students just below the cut point was $-.15$ (the difference of $-.28$). In comparison, for the 21 low-achieving schools, classroom mean ability for students just above and below the cut point was, respectively, $-.28$ and $-.64$ with the difference of $-.36$. This suggests that for both students just below and above the cut score, classroom peer ability is affected by school academic composition; peer ability is substantially lower in low-achieving schools than high-achieving schools for both students below and above the cut score. However, the impact of scoring below the cut point on classroom peer ability is relatively similar between high- and low-achieving schools.

16. Burris, Hubert, and Levin (2006) examined the effect of accelerating math instruction for all students in mixed-ability classrooms based on suburban middle schools and found that this model did not reduce the learning of higher skill students. However, in this study, more than 90% of students had prior skills above the national average. Thus, the finding is unlikely to be generalizable to a context, such as Chicago, where many students have skills well below grade level.

References

Abdulkadiroğlu, A., Angrist, J., & Pathak, P. (2014). The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica*, *82*, 137–196.

Achieve. (2009). *Closing the expectations gap: Fourth annual 50-state progress report on the alignment of high school policies with the demands of college and careers*. Washington, DC: Author.

Allensworth, E., Nomi, T., Montgomery, N., & Lee, V. E. (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.

Ammermueller, A., & Pischke, J.-S. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, *27*, 315–348.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–472.

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

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

Barr, R., & Dreeben, R. (1983). *How schools work*. Chicago, IL: University of Chicago Press.

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Record*, *110*, 608–645.

Burris, C. C., Hubert, J. P., & Levin, H. M. (2006). Accelerating mathematics achievement using heterogeneous grouping. *American Educational Research Journal*, *43*, 105–136.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *1422*, 636–654.

Cortes, K. E., Goodman, J., & Nomi, T. (2015). Intensive math instruction and educational attainment: Long-run impacts of Double-Dose algebra. *Journal of Human Resources*, *50*, 108–158.

Council of the Great City Schools. (2009). *Urban indicator*. Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/domain/35/publication%20docs/Urban_Indicator09.pdf

de la Fuente, A., & Ciccone, A. (2002). *Human capital in a global and knowledge-based economy: Final report*. Brussels, Belgium: European Commission, Directorate-General for Employment and Social Affairs.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects and the impacts of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*, 1739–1774.

Duncan, G. J., Morris, P., & Rodrigues, C. (2011). Does money really matter? Estimating impacts of family income on young children's achievement with data from random-assignment experiments. *Developmental Psychology*, *47*, 1263–1279.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.

Gamoran, A. (2010). Tracking and inequality: New directions for research and practice. In M. Apple, S. J. Ball & L. A. Gandin (Eds.), *The Routledge international handbook of the sociology of education* (pp. 213–228). London, England: Routledge.

Gamoran, A., Nystrand, M., Berends, M., & LePore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal*, *32*, 678–715.

Gibbs, C. (2015). *Experimental evidence on early intervention: The impact of full-day kindergarten*. Unpublished manuscript.

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

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). *Empirical benchmarks for interpreting effect sizes in research*. New York, NY: MDRC.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, *14*, 205–227.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association*, *101*, 901–910.

Hoxby, C. M., & Weingarth, G. (2005). *Taking race out of the equation: School reassignment and*

the structure of peer effects. Unpublished manuscript.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635.

Imberman, S., Kugler, A. D., & Sacerdote, B. (2012). Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *American Economic Review*, *102*, 2048–2082.

Kelly, S. (2004). Are teachers tracked? On what basis and with what consequences. *Social Psychology of Education*, *7*, 55–72.

Kelly, S. (2008). What types of students' effort are rewarded with high marks? *Sociology of Education*, *81*, 32–52.

Kemple, J. J., Herlihy, C. M., & Smith, T. J. (2005). *Making progress toward graduation: Evidence from the talent development high school model*. New York, NY: MDRC.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*, 83–119.

Lavy, V., Silva, O., & Weinhardt, F. (2012). The good, the bad, and the average: Evidence on ability peer effects in schools. *Journal of Labor Economics*, *30*, 367–414.

Loveless, T. (2009). *Tracking and detracking: High achievers in Massachusetts middle schools*. Washington, DC: The Thomas B. Fordham Institute.

Mac Iver, D. J., Balfanz, R., & Plank, S. B. (1998). An "elective replacement" approach to providing extra help in math: The Talent Development Middle Schools' Computer- and Team-Assisted Mathematics Acceleration (CATAMA) Program. *Research in Middle Level Education Quarterly*, *22*(2), 1–23.

Murnane, R. J., Willett, J. B., & Levy, F. (1994). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics*, *77*, 251–266.

National Governors Association Center for Best Practices. (2005). *Getting it done: Ten steps to a state action agenda*. Retrieved from http://www.nga.org/files/live/sites/NGA/files/pdf/05WARNERGUIDE.pdf

Nomi, T. (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, T., & Allensworth, E. (2009). "Double-Dose" algebra as an alternative strategy to remediation: Effects on students' academic outcomes. *Journal of Research on Educational Effectiveness*, *2*, 111–148.

Nomi, T., & Allensworth, E. (2013). Sorting and supporting: Why Double-Dose algebra led to better test scores but more course failure through changes in classroom composition, climate and instruction. *American Educational Research Journal*, *50*, 756–788.

Oakes, J. (2005). *Keeping track: How schools structure inequality* (2nd ed.). New Haven, CT: Yale University Press.

Page, R. N. (1991). *Lower track classroom: A curricular and cultural perspective*. New York, NY: Teachers College Press.

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, F. M. (1994). Ability-group effects: Instructional, social, or institutional? *Sociology of Education*, *67*, 27–46.

Park, H. (2013). *Re-examining education in Japan and Korea: Demystifying stereotypes*. New York, NY: Routledge.

Raudenbush, S. 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, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE.

Reardon, S., & Raudenbush, S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods and Research*, *42*, 143–163.

Rosenbaum, J. E. (1976). *Making inequality: The hidden curriculum of high school tracking*. New York, NY: Wiley.

Rosenbaum, J. E. (1999). If tracking is bad, is detracking better? *American Educator*, *23*, 24–29, 47.

Rubin, B. C. (2008). Detracking in context: How local constructions of ability complicate equity. *Teachers College Record*, *110*, 646–699.

Rubin, D. B. (1986). Which ifs have causal answers? *Journal of the American Statistical Association*, *81*, 961–962.

Sampson, R. (2012). *Great American city: Chicago and the enduring neighborhood effect*. Chicago, IL: University of Chicago Press.

Schafer, E. W., & Olexa, C. (1971). *Tracking and opportunity: The locking-out process and beyond*. Scranton, PA: Chandler.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*, 309–317.

Wheelock, A. (1992). *Crossing the tracks: How "untracking" can save America's schools*. New York, NY: New Press.

White, P. A., Gamoran, A., Smithson, J., & Porter, A. C. (1996). Upgrading the high school math curriculum: Math course-taking patterns in seven high schools in California and New York. *Education Evaluation and Policy Analysis*, *18*, 285–307.

Authors

TAKAKO NOMI is an assistant professor of education at Saint Louis University. Her areas of specialization include urban education, education policy, inequality in education, social organization of schools, and research methodology. Her recent work examines how district or state education policies affect the distribution of academic outcomes, variation in the intervention effects, and mechanisms to account for impact variation. She received a PhD in educational policy and theory from Penn State University.

STEPHEN W. RAUDENBUSH is the Lewis-Sebring Distinguished Service Professor in the Department of Sociology, the College and the Harris School of Public Policy, and chairman of the Committee on Education at the University of Chicago. He is interested in statistical models for child and youth development within social settings such as classrooms, schools, and neighborhoods. He is best known for his work developing hierarchical linear models, with broad applications in the design and analysis of longitudinal and multilevel research. He is currently studying the development of literacy and math skills in early childhood with implications for instruction, and methods for assessing school and classroom quality. He is a member of the National Academy of Sciences, the American Academy of Arts and Sciences, and the recipient of the American Educational Research Association award for Distinguished Contributions to Educational Research. He was awarded an EdM in administration, planning, and social policy, and an EdD in policy analysis and statistics from Harvard University.