AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–15
DOI: 10.1177/2332858420979167
Article reuse guidelines: sagepub.com/journals-permissions

Schools With Test-Based Promotion: Effects on Instructional Time Allocation and Student Learning in Grade 3

Yihua Hong
RTI International

Guanglei Hong
The University of Chicago

This study is focused on the threat of retention associated with test-based promotion in Grade 3. Through analyzing the Early Childhood Longitudinal Study Kindergarten Class of 1998–1999 data, we found that schools having such a policy apparently increased math instructional time but not reading instructional time in Grade 3. On average, the policy did not produce significant differences in third graders' reading and math learning. However, there seemed to be a notable increase in the proportion of students who achieved an at or above-average proficiency level in Grade 3 math. In both reading and math, the test-based promotion seemingly benefited students at the average or lower than average ability levels. In contrast, there was no evidence that the policy had an impact on students at the two ends of the ability distribution. We discussed the implication of the findings for the current design and implementation of test-based promotion in early grades.

Keywords: *propensity score, multilevel modeling, test-based accountability*

Since the late 1990s, a number of states (e.g., Florida, Georgia, Louisiana, North Carolina, and Texas) and large school districts (e.g., Chicago, New York City) have enacted promotional gates and required students to score above a certain minimum level on standardized tests, typically in reading and/or math, before being promoted to the next grade (Huddleston, 2014). At present, this practice also known as test-based promotion has continued to receive endorsement from policy makers. A particular emphasis has been placed on reading proficiency in third grade. For example, 16 states plus D.C. have mandated retention for third graders who do not demonstrate proficiency on state reading tests; eight extra states have similar practices at the discretion of local schools or school districts (National Conference of State Legislatures [NCSL], 2019). Proponents of test-based promotion have argued that students would likely respond to the promotion standard by working harder and that retained students could benefit from an additional year of schooling. However, opponents claimed that such a policy would violate professional standards for fair and appropriate test use and that retained students might suffer from reduced expectations of teachers and parents, difficulties in adjusting to a younger peer group, and an increased likelihood of school dropout (see reviews of Huddleston, 2014 and Penfield, 2010).

Past research on test-based promotion often focused on the effectiveness of grade repetition and mostly attended to retained students and their performance in post-retention years (e.g., Eren et al., 2017; Roderick & Nagaoka, 2005; Schwerdt et al., 2017; Winters & Greene, 2012). However, in addition to holding back low-performing students, test-based promotion is also expected to exert influence on teaching and learning through the threat of retention. For example, it may direct more attention and resources to low-preforming students prior to or at promotional gate grades; and such students may exert more learning efforts under pressure (Thomas, 2005). Research on the Ending Social Promotion reforms in New York city and Chicago (Allensworth & Nagaoka, 2010; McCombs et al., 2009) suggested that introducing the policy of test-based promotion may change teacher behavior, increase the amount of time allocated to core subjects or skills, and shift resources toward test preparation. It was accompanied with gains in reading and math test scores in both promotional and nonpromotional gate grades. Moreover, the policy seemed to have differential effects on student test performance with benefits particularly for students at lower than average ability levels or students with moderate or high risk of retention. However, the effects on teaching and learning were mostly observed in upper grades. In addition, because high-stakes consequences were

attached to the test scores that were also used as study outcomes, the observed score gains might not reflect actual improvement in student learning (B. A. Jacob, 2003, 2005). Limited to two large urban school districts, the generalizability of the findings remains uncertain.

The goal of this study is to examine how early-grade teachers and students responded to the threat of retention when the test-based promotion policy was present in schools. Building on the existing research, we analyze the Early Childhood Longitudinal Study Kindergarten (ECLS-K) and use a propensity score-based approach, the marginal mean weighting through stratification (MMW-S; Hong, 2010, 2012, 2015) to compare schools with and without test-based promotion in the amount of time teachers spent on the core subjects in Grade 3 (i.e., reading and math) and in students' learning in these subjects as measured by the ECLS direct assessment. We ask the following questions: (a) Did the presence of test-based promotion increase the amount of time allocated to reading and math instruction? (b) Did it affect student math and reading learning? and (c) Did the effects on the learning differ by student prior academic ability levels?

Although ECLS has collected data from two kindergarten cohorts, we use the data from the first cohort (1998–1999) in year 2002 which corresponds to the beginning of the No Child Left Behind (NCLB) era. At the time, the test-based promotion policy had been introduced in a few places (e.g., Louisiana and Chicago) and started to roll out at scale in some other states and large school districts (e.g., Florida, Georgia, Texas, Wisconsin, and New York City, see review of Huddleston, 2014). A battery of items measuring school retention practices were included in the ECLS school administrator questionnaire. Despite variations across states and districts and changes in promotion standards and practices over time (Huddleston, 2014; Marsh et al., 2009; NCSL, 2019), Grade 3 has generally been considered as an early check point for student progress and a critical milestone of student learning especially in terms of reading skills (Hernandez, 2011). By focusing on all students and their teachers in Grade 3 in the nationally representative ECLS sample and examining student performance on low-stake assessments that are not prone to the problem of score inflation due to inappropriate test preparation, this study adds to a small body of literature that examines the effects of retention threat associated with test-based promotion. Specifically, we aim to provide a historical picture of whether and how the policy exerted its influence on teaching and learning in early grades.

Teachers' Behavioral Changes Under Test-Based Promotion

Test-based promotion holds students accountable for their own performance and imposes no explicit sanctions or rewards for teachers. However, as a teacher's success is often defined by student performance (Cohen, 1996; Finnigan & Gross, 2007; Johnson, 1986), test-based promotion may potentially change teachers' instructional decision making (Allensworth & Nagaoka, 2010). Past literature on standards-based reforms has suggested that the standards used in test-driven accountability represent instructional targets for teachers and may influence the content and emphasis of instruction. Researchers (e.g., Au, 2007; Koretz et al., 1996; Taylor et al., 2001) have found that while investing more time in instructional alignment with grade-specific standards, in order to help more students meet the testing standards, teachers may significantly narrow the curriculum and focus less on intellectually challenging work. Polikoff (2012) showed that, once the NCLB was in effect, the introduction of standards and assessments was associated with an improvement in instructional alignment with state standardized tests, producing the largest and the most consistent increases in the alignment in mathematics across all grades.

Similarly, in a study of the Ending Social Promotion reform in Chicago, R. T. Jacob et al. (2004) found that after the school district adopted test-based promotion, teachers invested significantly more time in teaching math, especially in teaching grade-level math skills in the seventh and eighth grades. The increase in the math instructional time was over one half of a standard deviation above the prepolicy level. A small yet insignificant upward trend was observed for teaching reading comprehension over the postpolicy years. In addition, teachers tended to provide more instructional support to low-achieving student after the introduction of the Chicago Public Schools test-based promotion policy—the changes were mostly found in upper grades. In Grade 3, the researchers only observed a decline in introducing new math topics. Diamond's (2007) observation of second- and fifth-grade classrooms revealed that while the Chicago policy affected teacher behaviors, the changes seemed to be limited to the alignment with tests content and format without adding to the complexity and depth of instruction. Summarizing the past evidence, we suspect that test-based promotion may exert influence on teaching mainly through shifting instructional time and mobilizing resources to concentrate on tested subjects and grade-level skills.

Third-Grade Learning Under Test-Based Promotion

Evidence from qualitative studies has suggested that test-based promotion may affect student learning through changing student motivation (e.g., Roderick & Engel, 2001). However, this may not be the case with primary graders who tend to make little differentiation between ability and effort and tend to be optimistic about one's own academic competency (Dweck, 2001; Stipek, 2002). Even in Grade 3, rather than directly motivating students, test-based promotion is more likely to affect student learning indirectly through

changing teachers' instructional practices that in turn shape students' learning opportunities at school (Nye et al., 2004; Palardy & Rumberger, 2008; Rowan et al., 1997).

Teachers' behavioral changes under test-based promotion, as Allensworth and Nagaoka (2010) pointed out, could be a double-edged sword for student learning. While students at-risk of retention may benefit from increased attention from teachers, high-achieving students who are at no risk of retention are unlikely to receive a similar level of support (R. T. Jacob et al., 2004). In addition, instructional alignment with promotion standards may create differential learning patterns among students at different ability levels. Past studies on reading or math learning in early grades have shown that students with low academic skills benefit from the exposure to basic content while their peers with high academic skills benefit from more advanced content (Engel et al., 2013; Wonder-McDowell et al., 2011; Xue & Meisels 2004). It is likely that aligning instructional content with promotion standards that emphasize the mastery of grade-level skills may be beneficial only to students whose academic proficiency is near the grade benchmark, but may not meet the learning needs of students who are far above or far below the grade proficiency level.

Existing studies on the relationship between test-based promotion and student academic learning prior to or at the promotional gate grades were mostly conducted in the context of the Ending Social Promotion reforms in New York City (NYC) or Chicago. Researchers of NYC's policy (Mariano et al., 2009) compared all fifth graders who were in the promotional gate year to a same-grade cohort who had not been exposed to the policy and found a significant improvement in students' test performance in English language arts and math. The researchers then grouped students into four proficiency levels with Level 1 as the lowest proficiency and 4 as the highest based on their fourth-grade assessments. Aiming to remove the potential confounding of concurring reform initiatives, they subtracted the observed gain in Level 3 students from that of Levels 1 and 2 students. The difference-in-differences strategy assumed that test-based promotion only affected students who were at risk of being retained—an assumption that would require an empirical verification. It was claimed that in English language arts, the average effect directly attributable to the test-based promotion was almost zero for students who were at the lower end of Level 1 but ranged from 0.10 to 0.21 standard deviations for Level 2 and other Level 1 students who were close to the retention cutoff; in math, the average adjusted effect was indistinguishable from zero for all students at Levels 1 and 2.

Chicago Researchers similarly found that students experienced dramatic gains in test performance from third to eighth grades right after the introduction of the promotion policy, with the highest gains occurred at the promotional gate grades, that is, Grades 3, 6, and 8 (Bryk, 2003; B. A. Jacob, 2003, 2005; Roderick et al., 2002). By grouping students into different retention risk levels on the basis of the learning gain each student would need to achieve to reach the test cutoff, Roderick et al. (2002) revealed that for third graders, in reading, the threat of retention appeared to have a positive effect on the performance of high- or moderate-risk students but a negative effect on the performance of no-risk students; in math, it appeared to benefit all students with the largest effect on high- or moderate-risk students and the smallest effect on no risk students. B. A. Jacob (2003, 2005) argued that the observed test scores were unlikely a valid measure of student learning because the gains did not sustain over time despite the continuation of the policy initiatives and were not found on other low- or moderate-stakes tests. He pointed out that, as was the case in many studies on test-based accountability (e.g., Krieg, 2008; Ladd & Lauen, 2010; Neal & Schanzenbach 2010; Reback, 2008; Springer, 2008), standardized test scores were also used for promotion and other high-stakes decisions and were subject to the problem of score inflation by inappropriate test preparation. Hence, it remains a question whether the threat associated with test-based promotion can actually trigger large-scale changes in student learning and whether the distributional effects observed on high-stakes test scores in the previous studies can be similarly found on other low-stakes learning measures.

Method

Data and Measures

This study uses the first five waves of the ECLS-K data collected by the U.S. National Center for Education Statistics (NCES) from fall 1998 to spring 2002. The ECLS-K study sampled 18 students per school on average at kindergarten entry in fall 1998 and followed them over time. Spring 2002 was the time when most of the sampled students progressed toward the end of Grade 3, the grade that often serves as the lowest promotional gate grade. Standardized testing was prevalent during that year—only 7% of the sampled schools did not report having standardized testing in place. Such schools tended to have a relatively smaller enrollment, a larger proportion of disadvantaged students (e.g., limited English proficiency students, minority students, or students qualified for free or reduced-price lunch program), and a larger likelihood of providing gifted program services. This study focuses on 1,498 schools whose administrators reported having standardized testing in schools and provided information about practices of grade repetition. The total study sample includes 3,324 Grade–3 classrooms and 9,488 students.[1]

Test-Based Promotion. The treatment of interest in this study is the threat of retention associated with test-based promotion. According to school administrators' report in 2002, 278 schools (18.56%) reported that students could be retained if they failed a school-wide standardized test. These

are considered as the treatment group in this study; 1,220 schools (81.44%) did not report such a practice and thus constitute the comparison group.

Instructional Time. Our first research question is whether the presence of test-based promotion increased the amount of time allocated to reading and math instruction. We computed the amount of instructional time that third-grade teachers allocated to each subject based on the teacher self-reported frequency and duration of reading and math instruction in spring 2002. On average, the teachers spent about 385 minutes (SD = 134.80) per week teaching reading and about 269 minutes (SD = 100.82) teaching math.

Student Academic Learning. Our second research question asks whether the threat of test-based retention affected student learning of reading and math in Grade 3. The student learning was measured by the ECLS third grade direct assessments in spring 2002. The assessments were designed based on National Assessment of Educational Progress framework and consisted of test batteries in core subjects of each sampled grade (NCES, 2002). The test results, reported for each of these subjects, were not used by school districts or schools for accountability-related decision making. Hence unlike other assessment data analyzed in previous studies that were susceptible to score inflation due to attached high-stakes consequences, the ECLS assessment data promise to provide uncontaminated information about the effects of test-based promotion on student learning.

We used two types of ECLS assessment scores to measure student achievement at the end of Grade 3. One is the *standardized T scores* in reading and math that are norm-referenced with a mean of 50 and a standard deviation of 10 for the entire ECLS sample. These scores measure a student's knowledge and skills in each of these two subjects in comparison with other students in the same cohort. The other type of scores is *proficiency levels* in reading and math, a criterion-referenced measure that compares a student's knowledge and skills to the predetermined standards or performance levels. If a student demonstrated the required skills and knowledge at one proficiency level, they should have passed lower levels as well (NCES, 2005). According to the National Institute of Child Health and Human Development (2000), the focal emphasis of a Grade 3 reading curriculum is on comprehension skills, including making inferences with cues that are directly stated with key words in text and identifying clues used to make inferences. The National Council of Teachers of Mathematics (2009) stated that a Grade 3 math curriculum should focus on gaining fluency in using all basic multiplication and division facts and that efficient use and in-depth understanding of the operations also require conceptual knowledge of place value. Using these grade-level benchmarks, we combined the proficiency levels of each subject into four categories: 1 = *below-grade proficiency*; 2 = *Grade 3 below-average proficiency*; 3 = *Grade 3 at- or above-average proficiency*; 4 = *above-grade proficiency* (see Table 1).

Student Prior Academic Ability. Our third research question asks whether the threat of test-based retention affected students differently depending on their prior academic ability. The ECLS-K 98 study did not survey or assess students in Grades 2 and 4. Although we have information about the proportion of students in a Grade–3 class who were already repeating the grade in spring 2002, no data were collected on the promotion standards adopted by schools or school districts or on student retention status in Grades 3 and 4. As a result, we were unable to determine a student's risk of being retained at the end of Grade 3. Following the studies in NYC and Chicago (Allensworth & Nagaoka, 2010; McCombs et al., 2009), we decided to use a student's predicted prior reading ability at the end of Grade 2 as a moderator. The ECLS data contain repeated measures of student reading performance in the falls and springs of Kindergarten and Grade 1 and in the fall of Grade 3 that were vertically equated on the same metric. Using an empirical Bayes estimation approach, we estimated a nonlinear reading growth trajectory in kindergarten and Grade 1 for each student and extrapolated it to the end of Grade 2 under the assumption that a student would likely stay on the same growth trend prior to Grade 3 (see online Supplemental Material A). Based on the distribution of the predicted score, we classified the sampled students into five equal-sized ability groups, with 1 being the lowest ability group and 5 the highest group.

Pretreatment Covariates. The ECLS-K study did not collect information on the exact year when the test-based promotion was first introduced into each treated school. We found that the proportion of students who were already repeating Grade 3 in spring 2002 was twice as many in the treatment schools (M = 3.54, SD = 4.25) as in the comparison schools (M = 1.79, SD = 3.27) though the treated schools and the comparison schools did not differ significantly in the proportion of students at or above grade level in reading or verbal skills by the end of Grade 3. Such evidence might indicate that some of the treated schools had adopted the test-based promotion policy prior to 2002. If schools started implementing the policy earlier than fall 1998, the beginning of the ECLS-K study, certain school and student characteristics observed between kindergarten and Grade 3 (e.g., school climate and student reading and math assessment scores in spring kindergarten and in fall and spring Grade 1) might have been altered by the test-based promotion and thus should be considered as posttreatment rather than pretreatment variables. If so, controlling for these variables would likely lead to an underestimation of the policy effects. Therefore, we selected only two types of pretreatment covariates that are unlikely

TABLE 1
Reading and Math Proficiency Levels

Label	Original ECLS proficiency levels Skill content
Reading	
Below-grade proficiency	1. Letter recognition: identifying upper- and lower-case letters by name
	2. Beginning sounds: associating letters with sounds at the beginning of words
	3. Ending sounds: associating letters with sounds at the end of words
	4. Sight words: recognizing common words by sight
	5. Comprehension of words in context
Grade 3 below-average proficiency	6. Literacy inference: making inferences using cues that were directly stated with key words in text
Grade 3 at- or above-average proficiency	7. Extrapolation: identifying clues used to make inferences
Above-grade proficiency	8. Evaluation: demonstrating understanding of author's craft and making connections between problem in the narrative and similar life problems
Math	
Below-grade proficiency	1. Number and shape: identifying some one-digit numerals, recognizing geometric shapes, and one-to-one-counting of up to 10 subjects
	2. Relative size: reading all single-digit numerals, counting beyond ten, recognizing a sequence of patterns, and using nonstandard units of length to compare subjects
	3. Ordinality, sequence: reading two-digit numerals, recognizing the next number in a sequence, identifying the ordinal position of an object, and solving a simple word problem
	4. Addition/subtraction: solving simple addition and subtraction problems
Grade 3 below-average proficiency	5. Multiplication/division: solving simple multiplication and division problems and recognizing more complex number patterns
Grade 3 at- or above-average proficiency	6. Place value: demonstrating understanding of place value in integers to hundreds place
Above-grade proficiency	7. Rate and measurement: using knowledge of measurement and rate to solve word problems
	8. Fractions: demonstrating understanding of the concept of fractional parts

Note. ECLS = Early Childhood Longitudinal Study.

the results of the test-based promotion yet may predict instructional time or student learning: (a) demographic features of students, classes, and schools that are relatively stable overtime and (b) students' baseline academic and social emotional status as well as their previous learning experience, which were measured at the time of kindergarten entry (see online Supplemental Material B for the list of pretreatment variables). To impute the missing information in the selected variables, we used a multistage imputation procedure that took into account the multilevel structure of the data.[2] Because the treatment assignment was a school-level measure, all 81 pretreatment covariates were then aggregated to the school level.

Analytic Strategies

Corresponding to the research questions, we conducted three sets of quasi-experimental analyses. The first set of analyses examined the overall effects of test-based promotion on instructional time allocation; the second set considered the effects of the threat of test-based retention on student academic learning; and the third set investigated whether the effects on learning depend on students' prior ability levels. Since the assignment to the two treatment conditions was unlikely to be random, we employed a semiparametric propensity score-based strategy—MMW-S (Hong, 2010, 2012, 2015)—to reduce selection bias associated with the rich set of pretreatment covariates in the ECLS-K data. MMW-S has shown promises for evaluating various types of treatments and for investigating moderated treatment effects in educational studies (Garrett & Hong, 2016; Hong et al., 2012; Hong & Hong, 2009). By assigning a weight to each unit based on propensity stratification, the pretreatment composition of a weighted treatment group resembles that of the whole population or a sub-population.

Effects on Instructional Time Allocation. For the first set of analysis, we used MMW-S to approximate a simple randomized design in which schools were "as if" randomly assigned to either the "test-based promotion" group or the comparison group. We estimated each school's propensity for adopting test-based promotion through analyzing a logistic regression

model. A school's propensity score is the estimated conditional probability as a function of the 81 school-level covariates. Based on the propensity score, we then subdivided the school sample into four strata and computed a weight for each treated school ($z = 1$) and each comparison school ($z = 0$) in stratum s as a ratio of the *expected* number of schools to the *actual* number of schools assigned to treatment condition z in that stratum. Here the numerator is the product of the total number of schools in stratum s (denoted by n_s) and the proportion of schools assigned to treatment condition z (i.e., $Pr\{Z = z\}$). The denominator is the actual number of treated or comparison schools in that stratum (denoted by $n_{z,s}$):

$$MMW = \frac{n_s}{n_{z,s}} \times Pr\{Z = z\} \tag{1}$$

After weighting, the two treatment groups were expected to become comparable in the distribution of the observed school-level propensity scores. A further analysis revealed that the two weighted treatment groups no longer differed in the distribution of about 96% of the school-level covariates.

To estimate the effects on instructional time allocated to reading and math, we ran a weighted two-level multivariate model with teachers nested within schools and with the marginal mean weight applied at the school level. For teacher j in school k,

$$\begin{aligned} Time_{jk} &= Dread_{jk} * (\beta_0 + \beta_1 Z_k) + Dmath_{jk} * (\beta_2 + \beta_3 Z_k) \\ &+ \boldsymbol{\beta_T T_{jk}} + \boldsymbol{\beta_s S_k} + u_k + r_{jk},\ u_k \sim N(0, \tau_u), r_{jk} \sim N(0, \tau_r), \end{aligned} \tag{2}$$

The outcome "*Time*" is the instructional time spent on reading or math. Here, we used dummy indicators "*Dread*" and "*Dmath*" to indicate whether the outcome was for reading or math, respectively. To improve precision and reduce residual bias, we additionally controlled for a vector of class-level outcome predictors *T*, which included proportion of non-English speakers and proportion of students eligible for free or reduced price lunch/breakfast, and a vector of important school-level outcome predictors *S* including school region, low-income school, an aggregated measure of current students' approaches to learning rated in fall 1998, proportion of students below 9 years old, and proportion of white students. In the absence of unmeasured confounders, β_1 and β_3 were used to identify the effects of test-based promotion on instructional time allocated to reading and math, respectively.

Effects on Student Learning. To examine the effects of test-based promotion on student learning, we used either standardized *T* scores or proficiency levels as the outcome. For the continuous standardized score (*Tscore*) of student i in the class of teacher j in school k, we ran a weighted three-level multivariate model with the same school-level weight:

$$\begin{aligned} Tscore_{ijk} &= Dread_{ijk} * (\beta_0 + \beta_1 Z_k) + Dmath_{ijk} * (\beta_2 + \beta_3 Z_k) \\ &+ \beta_4 S_k + \beta_5 T_{jk} + \boldsymbol{\beta_x X_{ijk}} + u_k + r_{jk} \\ &+ e_{ijk}, u_k \sim N(0, \tau_u),\ r_{jk} \sim N(0, \tau_r),\ e_{ijk} \sim N(0, \sigma^2), \end{aligned} \tag{3}$$

We additionally adjusted for a vector of student outcome predictors *X* at Level 1 including age at kindergarten entry, socioeconomic status, kindergarten reading and math scores, and teacher-rated approaches to learning. We controlled for class proportion of white *T* at Level 2 and school region *S* (rural vs. the rest) at Level 3.

When examining the proficiency level as an outcome, we ran a weighted multinomial logit model for each subject (i.e., reading or math), the structural model is specified as

$$\begin{aligned} \eta_{ijk} &= \beta_0 + \beta_1 Z_k + \beta_2 S_k + \beta_3 T_{jk} + \boldsymbol{\beta_x X_{ijk}} \\ &+ u_k + r_{jk}, u_k \sim N(0, \tau_u),\ r_{jk} \sim N(0, \tau_r), \end{aligned} \tag{4}$$

With the Grade 3 below-average proficiency as the reference level, η is the log odds of achieving each of the rest three proficiency levels (i.e., below-grade proficiency, Grade–3 at- or above-average proficiency, or above-grade proficiency) to the reference level. We controlled the same set of covariates as in Equation 3.

Effects on Academic Performance by Student Ability Levels. In the last set of the analyses, we investigated whether the treatment effects were differential by student prior ability levels. Here, we intended to approximate a block randomized design in which five subpopulations of students defined by their prior ability level were viewed as blocks. Within each subpopulation of students, the schools that they attended in Grade 3 were "as if" randomized to either test-based promotion or the comparison condition. On the basis of the four school-level strata previously obtained, for a student from subpopulation a whose school was in treatment group z and in stratum s, the marginal mean weight is:

$$MMW2 = \frac{n_{(s|A=a)}}{n_{(z,s|A=a)}} \times Pr\{Z = z \mid A = a\} \tag{5}$$

We applied the student-level weight to a three-level univariate model for estimating the moderated effects of test-based promotion on student learning in each of the two subjects as measured by the standardized *T* scores:

$$\begin{aligned} TScore_{ijk} &= ability1(\beta_0 + \beta_1 Z_k) + ability2(\beta_2 + \beta_3 Z_k) \\ &+ ability3(\beta_4 + \beta_5 Z_k) + ability4(\beta_6 + \beta_7 Z_k) \\ &+ ability5(\beta_8 + \beta_9 Z_k) + \beta_{10} S_k + \beta_{11} T_{jk} \\ &+ \boldsymbol{\beta_x X_{ijk}} + u_k + r_{jk} + e_{ijk},\ u_k \sim N(0, \tau_u),\ r_{jk} \\ &\sim N(0, \tau_r),\ e_{ijk} \sim N(0, \sigma^2), \end{aligned} \tag{6}$$

TABLE 2
Comparison Between the ECLS-K-3 Full Sample and the Analytic Sample

Parameter	Full sample	Analytic sample
Number of students	17,401	8,200
Number of classes	5,968	3,101
Number of schools	2,954	1,329
Proportion of public schools	0.8	0.87***
Proportion of schools in Northeast	0.19	0.15***
Proportion of schools in Midwest	0.26	0.24***
Proportion of schools in South	0.33	0.36***
Proportion of schools in West	0.23	0.25***
Proportion of urban schools	0.38	0.39**
Proportion of suburban schools	0.4	0.4
Proportion of rural schools	0.22	0.21***
Proportion of schools with 6th and above grades	0.25	0.20***
Proportion of small-size schools (less than 150 students)	0.05	0.02***
Average age at 3rd grade entry	8.46	8.48***
proportion of girls	0.49	0.51***
Proportion of White students	0.69	0.7
Proportion of Black students	0.12	0.11*
Proportion of Hispanic students	0.19	0.19
Proportion of Asian students	0.08	0.08
Proportion below poverty	0.19	0.18**
Proportion of students speaking non-English at home	0.13	0.15***

Note. ECLS-K = Early Childhood Longitudinal Study–Kindergarten. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. These indicate a statistically significant difference between the two samples.

Here, *ability*1, *ability*2, *ability*3, *ability*4, and *ability*5 are dummy indicators corresponding to the subpopulations of students at the five prior ability levels. This model estimates five treatment effects, one for each of these five subpopulations of students. These treatment effects are denoted by β_1, β_3, β_5, β_7, and β_9.

To allow for meaningful comparisons, every school in the analyses should have a nonzero probability of adopting or not adopting test-based promotion. For this reason, we excluded 169 schools that did not have any counterparts in the alternative treatment group on the basis of their estimated propensity scores. Our final analytic sample included 8,200 students from 3,101 classes in 1,329 schools. Table 2 compares the analytic sample with the ECLS-K-3 full sample (n = 17,401). The former included more public schools, schools from south or west regions, urban schools, and schools with Grade 6 and above. Relatively speaking, the schools in the analytic sample tended to be smaller in size, had a larger proportion of girls, a larger proportion of students speaking a non-English language at home, a smaller proportion of black students, and a smaller proportion of students below poverty.

Results

We started our analysis by first looking at school characteristics that predicted the selection into test-based promotion. With adjustment for the observed school-level pretreatment covariates, we then compared the schools with and without the threat of retention in terms of instructional time allocated to reading and math and in students' academic learning. We further investigated whether the test-based promotion exerted differential influences on students at different ability levels.

Characteristics of Schools With Test-Based Promotion

We found that the practice of retaining students based on standardized test results was more prevalent in suburban schools and schools in West or Midwest. It was more often found in private or catholic schools at that time. These schools usually enrolled sixth graders and above, and hence had a larger student enrollment and a larger proportion of regular teachers. In addition, compared with schools without test-based promotion, schools with the policy in place had a higher concentration of minority students (especially black and Hispanic students), students eligible for the reduced-price or free lunch program, students with limited English proficiency, and students with lower general knowledge scores at kindergarten entry. The treated schools tended to suffer more from limited educational resources (e.g., inadequate space and no service for students with special needs), problems of student mobility, and/or serious violence and crime within schools and around. Hence, we reason that, had the test-based promotion policy been absent in the treated schools, their students would likely have demonstrated reading and math performance at a lower level on average than their counterparts in the comparison schools.

Effects of Test-Based Promotion on Instructional Time Allocation

We first evaluated the effects of test-based promotion on the time allocated to reading and math instruction. The second panel of Table 3 summarizes the distribution of instructional time by treatment conditions. Grade 3 teachers allocated more instructional time to reading than math on average, a pattern that appeared to be consistent across the two treatment conditions. With adjustment for the observed pretreatment covariates, results from analyzing Model 2 suggest that the presence of test-based promotion in school did not produce a statistically significant change in the amount of time allocated to reading instruction. However,

TABLE 3
Outcomes Distribution by Treatment Conditions

Parameter	Comparison	Test-based promotion	Total
Number of schools	1,054	275	1,329
Number of classes	2,418	683	3,101
Number of students	6,599	1,601	8,200
Instructional time			
Reading	380.70 (134.06)	388.41 (138.01)	385.49 (134.80)
Math	264.37 (96.28)	284.89 (119.22)	268.63 (100.82)
Student standardized *t* score			
Reading	51.47 (9.08)	51.41 (8.95)	51.46 (9.05)
Math	51.55 (9.23)	52.04 (9.23)	51.65 (9.23)
Student proficiency level (%)			
Reading			
Below-grade proficiency	22.7	21.9	22.5
Grade 3 below-average proficiency	25.6	27.2	25.9
Grade 3 at- or above-average proficiency	28.4	29.0	28.5
Above-grade proficiency	23.3	22.0	23.0
Math			
Below-grade proficiency	20.4	19.7	20.2
Grade 3 below-average proficiency	31.4	30.2	31.1
Grade 3 at- or above-average proficiency	31.6	32.6	31.8
Above-grade proficiency	16.7	17.6	16.9

having such a policy apparently increased math instructional time by 15.13 minutes per week ($SE = 6.38$, $t = 2.37$, $p < .01$), which amounts to about 15% of a standard deviation.

Effects of Test-Based Promotion on Student Learning

To evaluate the policy effects on student learning in reading and math, we examined both standardized *T* scores and proficiency levels as outcomes.

The *standardized T scores* measured a student's academic performance relative to other students in the same cohort. Analyzing Model 3, we found no indication that test-based promotion led to a notable difference in students' reading or math performance on average between the two treatment conditions (reading: coefficient = 0.31, $SE = 0.35$, $t = 0.89$, $p > .05$; math: coefficient = 0.38, $SE = 0.37$, $t = 1.04$, $p > .05$). The effect size was no larger than 0.04 in both subjects.

The *proficiency levels* evaluated a student's academic performance against grade-level benchmarks. As shown in the bottom panel of Table 3, within each treatment condition, relatively fewer students were either below-grade or above-grade in their reading and math proficiency levels. Table 4 presents the results from analyzing the three-level multinomial logit model (Equation 4) using Grade 3 below-average proficiency as a reference level. In reading, we did not find any statistical evidence that having test-based promotion in school would change the likelihood that a student reaches a proficiency level different from the reference level. In math, however, test-based promotion appeared to be associated with a 25% increase in the odds of achieving Grade 3 *at- or above-average proficiency to below-average* proficiency (coefficient = 0.22, $SE = 0.10$, $t = 2.26$, $p < .05$).

Effects of Test-Based Promotion on Student Learning by Ability Levels

We hypothesized that the effects of test-based promotion might depend on whether a student was near the margin of failing the grade because such a student might receive special attention from the teacher. The analysis of the overall average effects of test-based promotion on the standardized *T* scores would likely leave such subpopulation-specific treatment effects undetected. Therefore, our subsequent analysis examined the differential effects by students' prior academic ability levels. Table 5 shows how student-standardized *T* scores distributed across treatment conditions and across ability levels. In analyzing Model 6, we first conducted omnibus tests comparing a model with the indicator for test-based promotion and a model without the indicator. The test results were statistically significant for both reading ($\chi^2_5 = 18.86, p < .01$) and math ($\chi^2_5 = 12.54, p < .05$), suggesting the existence of a significant effect of test-based promotion within at least one ability group for each subject.

We then proceeded to examine the treatment effects by ability levels. As shown in Table 6, the presence of test-based promotion had no effects on the average standardized

TABLE 4
Estimated Effects of Test-Based Promotion on Student Proficiency Levels

	Reading				Math			
	Fixed effects							
Proficiency level	Coefficient	Odds ratio	*SE*	*t*	Coefficient	Odds ratio	*SE*	*t*
For below-grade proficiency								
Intercept	−0.40	0.67	0.04	−9.04***	−0.64	0.53	0.04	−15.15***
Test-based promotion	−0.17	0.85	0.10	−1.63	0.11	1.11	0.12	0.92
Age at K entry	0.02	1.02	0.01	2.12*	0.00	1.00	0.01	0.29
SES	−0.31	0.73	0.05	−5.87***	−0.28	0.75	0.05	−5.25***
Approaches to learning rated by K teacher	−0.40	0.67	0.06	−7.07***	−0.51	0.60	0.06	−8.90***
K reading score	−0.01	0.99	0.00	−3.41***	0.01	1.01	0.00	1.97*
K math score	−0.00	1.00	0.00	−0.03	−0.02	0.99	0.00	−4.52***
Class proportion of White	−0.00	1.00	0.00	−3.98***	−0.00	1.00	0.00	−2.78**
Rural school	0.31	1.37	0.11	3.02**	0.17	1.19	0.10	1.81
For Grade 3 at- or above-average proficiency								
Intercept	0.05	1.05	0.04	1.30	−0.04	0.96	0.04	−0.99
Test-based promotion	−0.07	0.93	0.10	−0.76	0.22	1.25	0.10	2.26*
Age at K entry	0.01	1.01	0.01	0.71	−0.02	0.98	0.01	−2.44*
SES	0.27	1.31	0.05	5.23***	0.25	1.28	0.05	4.92***
Approaches to learning rated by K teacher	0.34	1.40	0.06	5.32***	0.30	1.35	0.05	5.59***
K reading score	0.01	1.01	0.00	3.00**	−0.01	0.99	0.00	−2.01*
K math score	0.01	1.01	0.00	1.61	0.03	1.03	0.00	7.35***
Class proportion of White	0.00	1.00	0.00	4.26***	0.01	1.01	0.00	4.50***
Rural school	−0.04	0.96	0.09	−0.43	−0.26	0.77	0.09	−2.79**
For above-grade proficiency								
Intercept	−0.44	0.64	0.05	−8.81***	−1.08	0.34	0.06	−16.85***
Test-based promotion	−0.04	0.96	0.12	−0.37	0.23	1.26	0.14	1.68
Age at K entry	0.01	1.01	0.01	0.98	0.02	1.02	0.01	2.05*
SES	0.50	1.65	0.06	7.70***	0.69	1.99	0.06	10.76***
Approaches to learning rated by K teacher	0.74	2.10	0.07	11.22***	0.45	1.56	0.07	6.23***
K reading score	0.01	1.01	0.00	2.89**	−0.01	0.99	0.00	−3.62***
K math score	0.01	1.01	0.01	2.77*	0.06	1.06	0.01	9.38***
Class proportion of White	0.01	1.01	0.00	8.88***	0.01	1.01	0.00	5.85***
Rural school	−0.11	0.90	0.12	−0.37	−0.29	0.75	−0.13	−2.25*
	Random effects							
	Variance	*df*	χ^2		Variance	*df*	χ^2	
Level 2								
Below-grade proficiency	0.10	443	2856.06***		0.16	443	2920.43***	
Grade 3 at- or above-average proficiency	0.01	443	1890.70***		0.01	443	1828.63***	
Above-grade proficiency	0.00	443	1651.88***		0.03	443	1760.60***	
Level 3								
Grade 3 at- or above-average proficiency	0.22	1326	1360.57		0.30	1326	1559.56***	
Above-grade proficiency	0.36	1326	1486.04**		0.54	1326	1661.33***	

Note. Grade 3 below-average proficiency was used as a reference group. The variance of the Level-3 random effect for below-grade proficiency was found insignificant in the analyses of reading and math outcomes and thus was removed from the model. K = kindergarten; SES = socioeconomic status.
*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 5
Distribution of Reading and Math Scores by Treatment Conditions and Ability Levels

Treatment condition	Ability 1	Ability 2	Ability 3	Ability 4	Ability 5	All
Number of students						
Test-based promotion	315	334	345	311	296	1,601
Comparison	1,325	1,306	1,295	1,329	1,344	6,599
Reading						
Test-based promotion	42.84 (9.00)	48.27 (6.30)	52.62 (6.58)	55.14 (6.56)	58.74 (6.94)	51.41 (8.95)
Comparison	42.72 (8.84)	48.95 (7.16)	51.85 (6.75)	54.77 (6.42)	58.91 (6.75)	51.47 (9.08)
Math						
Test-based promotion	44.89 (8.97)	49.10 (7.87)	52.91 (7.90)	55.43 (7.34)	58.39 (7.70)	52.04 (9.23)
Comparison	44.52 (8.50)	48.95 (8.18)	51.52 (7.63)	54.33 (7.85)	58.31 (7.49)	51.55 (9.23)

TABLE 6
Estimated Effects of Test-Based Promotion on Student Learning by Prior Ability Levels

	Reading			Math		
Fixed effects	Coefficient	*SE*	*t*	Coefficient	*SE*	*t*
Ability 1						
Intercept	44.49	0.26	174.43***	46.32	0.24	190.39***
Test-based promotion	0.28	0.64	0.45	0.49	0.62	0.80
Ability 2						
Intercept	49.31	0.22	225.70***	49.23	0.23	208.21***
Test-based promotion	1.22	0.55	2.22*	1.59	0.65	2.44*
Ability 3						
Intercept	51.50	0.19	275.06***	51.12	0.22	234.03***
Test-based promotion	1.48	0.54	2.74**	0.81	0.70	1.17
Ability 4						
Intercept	53.81	0.19	286.69***	53.33	0.24	225.33***
Test-based promotion	−0.22	0.44	−0.50	0.36	0.54	0.167
Ability 5						
Intercept	57.30	0.20	288.11***	56.57	0.25	230.51***
Test-based promotion	−0.25	0.36	−0.69	−0.09	0.57	−0.16
Age at K entry	0.03	0.02	1.21	0.04	0.02	1.71
SES	1.86	0.12	15.66***	2.06	0.13	16.38***
Approaches to learning rated by K teacher	2.00	0.12	16.37***	2.00	0.14	14.21***
K reading score	0.02	0.01	2.40*	−0.08	0.01	−8.17***
K math score	0.06	0.01	5.59***	0.18	0.01	15.98***
Class proportion of White	0.03	0.00	9.71***	0.03	0.00	8.32***
Rural school	−0.84	0.29	−2.87**	−1.26	0.34	−3.68***
	Variance	*df*	χ^2	Variance	*df*	χ^2
Random effect						
Level 1, *e*	36.80			42.55		
Level 2, *r*	0.54	1771	1846.37	3.98	1771	2200.82***
Level 3, *u*	4.68	1328	2255.64***	5.12	1328	2038.32***

Note. K = kindergarten; SES = socioeconomic status.
*$p < .05$. **$p < .01$. ***$p < .001$.

T scores for third graders who were either at the bottom (Ability 1) or in the top two tiers of the ability distribution (Ability 4 and Ability 5). This was true in both reading and math. However, the presence of test-based promotion appeared to benefit students who were just above the lowest tier yet below the average (Ability 2). These students scored 1.22 points higher ($SE = 0.55$, $t = 2.22$, $p < .05$) in reading and 1.59 points higher ($SE = 0.65$, $t = 2.44$, $p < .05$) in math than their counterparts attending the comparison schools. The effect sizes were .13 and .17, respectively. The policy also appeared to benefit students in the middle tier of the ability distribution (Ability 3) in reading. Such students scored 1.48 points higher ($SE = 0.54$, $t = 2.74$, $p < .01$) than their counterparts in the comparison schools, with an effect size equal to .16.

Discussion and Conclusion

Analyzing a national dataset collected at the dawning of the NCLB era, this study contributes to the limited literature on the effects of threat of retention associated with the test-based promotion policy. We compared schools with and without retention consequences attached to standardized testing results and examined their teachers' instructional time allocation and the third graders' learning in reading and math. We unpacked the policy effects on learning through utilizing ECLS direct assessment results that were not used by the sampled schools and school districts for accountability-related decision making and hence were unlikely to be contaminated by inappropriate test preparation and school gaming strategies (B. A. Jacob, 2003, 2005). The results revealed the likely impacts on teaching and learning when schools set about to implement the new policy.

Summary of Findings

Effects on Instructional Time Allocation. To examine potential changes in teacher behavior, we focused on the time teachers allocated to reading and math instruction. Our results were consistent with the past findings from Chicago that showed an increased investment in math instruction in contrast with limited changes in reading instructional time after the introduction of test-based promotion (R. T. Jacob et al., 2004). We found that teachers in schools with the test-based promotion policy on average spent about 15 more minutes per week teaching math than their counterparts in the comparison schools; yet the policy did not appear to generate a difference in reading instructional time. The average impact of the policy on math instructional time amounted to approximately 6% of the average weekly math instructional time when test-based promotion was not in place ($M = 264$ minutes).

Subject differences in early grades are well documented in past literature (Spillane, 2005). Schools tend to spend a large amount of time on reading instruction but considerably less time on math in early grades as literacy is considered to be a subject that pervades all disciplines and has established dominance in the daily schedule of elementary schools; in contrast, math is conventionally treated as a stand-alone subject with secondary importance. This pattern continued into Grade 3, as shown in Table 3 that compares instructional time between reading and math. Our findings suggest that when there was an extra pressure to compare student performance against academic standards in both reading and math, teachers and schools might respond by increasing instructional resources for math while leaving the instructional time for reading intact.

Effects on Student Learning. We analyzed two types of ECLS assessment measures: standardized *T* scores and proficiency levels. Previous studies (e.g., Bryk, 2003; Mariano et al., 2009) relied mostly on high-stakes tests to measure student academic performance and often reported significant gains in test performance as a result of test-based promotion. Our analysis of the standardized *T* scores from the low-stakes ECLS assessments revealed that test-based promotion, on average, were unlikely to produce a great amount of improvement in student learning. Nevertheless, according to the results of our analysis of proficiency levels, although having the test-based promotion policy did not increase reading proficiency among students, it increased the likelihood that a student would move from just below to just at or above the average level of Grade 3 proficiency in math. To gain further understanding of how test-based promotion might affect students differently, we examined the treatment effects for different subpopulations of students defined by their prior academic ability. The results indicated that student response to the threat of retention differed by their prior ability. In both reading and math, the policy exerted no influence on students in either the bottom tier or the top two tiers of the ability distribution. These students were either nearly certain to be retained or had almost no risk of retention. The policy appeared to have improved the reading and math performance of students in the tier just below the average ability and therefore were likely at the margin of being retained. Interestingly, students whose prior ability was in the middle tier also appeared to benefit from test-based promotion in reading but not in math. It is possible that the academic standards were elevated to a greater extent in reading than in math and thus placed students with average prior ability around the retention threshold. The pattern of differential treatment effects was consistent with the findings from Chicago (Roderick et al., 2002) and seemed to reflect a shift of teachers' attention toward students who were near the margin of being retained yet had a good chance of promotion through some instructional help.

To sum up, our analyses have provided evidence suggesting some potential benefits of test-based promotion in the early days of the NCLB reform. Specifically, the policy might have promoted the mastery of Grade 3 math proficiency and

improved the learning of students at some but not a severe risk of retention. However, our results have also suggested that the effectiveness of the policy was likely limited in terms of influencing teaching and learning. Although under the test-based promotion policy, teachers might see a possibility of enhancing math instruction through increasing its instructional time, the time allocated to reading instruction had likely reached its maximum already. In addition, the policy did not seem to have met the learning needs of students at the two ends of the ability distribution and therefore failed to generate a large-scale improvement in student learning. These findings may have implications for the design and implementation of the current test-based promotion policy that emphasizes reading proficiency in early grades (NCSL, 2019). Test-based promotion alone is likely insufficient for bringing about meaningful changes in teaching and learning that would benefit all students, especially those who tend to suffer from the greatest learning disadvantage.

Study Limitations and Future Research Agenda

Like most secondary data analysis, our work is limited by the available data. Our measures of test-based promotion and instructional time were constructed based on the self-reports of school administrators and teachers and may contain an unknown amount of measurement errors. Additional data collection approaches such as classroom observation or reporting from multiple resources could be employed in the future to further validate the findings.

Moreover, although the ECLS-K 98 data allowed us to examine the outcomes of test-based promotion measured with low-stake assessments, the promotion policy introduced nearly 20 years ago might not be identical to those implemented in more recent years. In addition, no information was provided in ECLS-K about in which year the policy was first introduced into the sampled schools and in which grades promotion decisions were actually tied to student test performance in those schools. For example, we found indication in the data that some treated schools might have adopted test-based promotion even before NCLB started to roll out. Nevertheless, this study does not intend to make inference about the impacts of implementing test-based promotion in specific grades. Rather, we are interested in the threat of retention associated with test-based promotion in general. At a minimum, our empirical evidence has shown how third graders and their teachers would respond to the threat of retention when such a policy was present in schools.

We employed a propensity-score based casual inference approach—the MMW-S method—to reduce selection bias associated with 81 observed pretreatment covariates. The propensity score approach assumed that the treatment assignment was independent of unobserved confounders given the observed covariates (Rosenbaum & Rubin, 1983, 1984). This assumption may not hold as we were conservative in selecting pretreatment covariates due to lack of information about the starting time of test-based promotion in each school. As a result, we considered student/class/school demographic characteristics and students' learning and behaviors measured at the kindergarten entry only. The adjustment for this set of covariates may not be adequate for removing selection bias. For example, if the test-based promotion was coupled with other reform initiatives such as professional development that might have improved teacher behavior and student learning, we would have likely overestimated the policy effects.

We therefore assessed the extent to which our conclusion would likely be altered by the omission of potentially important unobserved confounders through a sensitivity analysis in which we made further adjustment for a hypothetical confounding effect. Assuming that a bias introduced by some unobserved confounders could not be more severe than that of the strongest observed confounder, we used the bias associated with the pretest scores measured at the kindergarten entry as the referent value for the hypothetical confounding (see procedure described in Hong & Hong, 2009). This sensitivity analysis was focused on the initial estimates that were statistically significant (i.e., the treatment effects on the reading performance of Ability 2 and Ability 3 students, as well as the treatment effect on the math performance of Ability 2 students, see Table 6). Given that the students enrolled in the treated schools tended to have a greater disadvantage and would likely display lower academic performance on average than those enrolled in the comparison schools in the absence of the policy, we reason that our initial results would likely be an underestimate rather than an overestimate of the policy benefit. In reading, after removing the hypothetical negative confounding, the adjusted estimates were 1.27 (95% CI: [0.19, 2.35]) for Ability 2 students and 1.51 (95% CI [0.46, 2.58]) for Ability 3 students. In math, the adjusted estimate was 2.01 (95% CI [0.74, 3.28]) for Ability 2 students. All the 95% confidence intervals were still above zero. Hence, our initial conclusion about the potential benefits of test-based promotion for selected subpopulations of students appeared to be robust to potential omission of confounding.

Although this study employed a national data set, our analytic sample excluded schools that did not provide information on either retention or testing practices as well as schools that did not have counterparts under the alternative treatment condition. The analytic sample also excluded students who had no K–3 assessment data. Table 2 has shown that the analytic sample differed from the ECLS K–3 full sample in school size and student composition. Hence readers should use caution when they attempt to generalize the findings to other populations.

Last, the current analysis did not examine whether the effects of test-based promotion differed across classrooms or teachers. Teachers may calibrate instruction according to their perception of students' learning potential (Cohen et al., 2003). Future research may examine whether the effects of

test-based promotion may depend on teachers' growth mindset. One may further investigate the medication process by asking whether test-based promotion may affect student learning though changing instructional practices.

Acknowledgments

This research was supported by a dissertation grant from the American Educational Research Association, which receives funds for its "AERA Grants Program" from the National Science Foundation under Grant #DRL-0941014 and a William T. Grant Foundation Supplementary Mentoring Award for Junior Scholars of Color. Additional support came from the Spencer Foundation and the Social Sciences and Humanities Research Council of Canada. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

ORCID iD

Yihua Hong https://orcid.org/0000-0002-7829-9378

Notes

1. About 1,363 schools in the ECLS sample did not report information on their retention practices or on the use of standardized testing and thus were also excluded from the analytic sample. Compared with other schools with information on testing and retention practices, these schools tended to be larger in size and/or had more students from low-income families. However, they did not differ significantly in the race/ethnicity composition of students and teachers.

2. We impute the missing values using a five-stage maximum likelihood-based imputation procedure: (1) conduct initial imputation of student-level variables X with their missing indicators and obtain imputed variables X^*; (2) aggregate X^* to teacher level and impute teacher level variables T with the aggregated $\bar{X}^*$ and obtain imputed variables T^*; (3) aggregate $\bar{X}^*$ and T^* to school level and impute school level variables W with the aggregated $\bar{X}^*$ and T^* and obtain imputed variables W^*; (4) impute T with W^* and obtain imputed variables T^{**}; (4) impute X with W^* and T^{**} and obtain imputed variables X^{**}; The final imputed dataset include variables X^{**}, T^{**} and W^*.

References

Allensworth, E., & Nagaoka, J. (2010). Issues in studying the effects of retaining students with high-stakes promotion tests: Findings from Chicago (pp. 327–341). In J. L. Meece, & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development*. Routledge.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Bryk, A. S. (2003). No child left behind, Chicago style. In P. E. Peterson, & M. R. West (Eds.), *No child left behind? The politics and practice of school accountability* (pp. 242–268). Brookings Institution Press.

Cohen, D. K. (1996). Rewarding teachers for student performance. In S. H. Fuhrman, & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 60–114). Jossey-Bass.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119–142. https://doi.org/10.3102/01623737025002119

Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education*, *80*(4), 285–313. https://doi.org/10.1177/003804070708000401

Dweck, C. S. (2001). The development of ability conceptions. In A. Wigfield, & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 57–88). Academic Press. https://doi.org/10.1016/B978-012750053-9/50005-X

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students with what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, *35*(2), 157–178. https://doi.org/10.3102/0162373712461850

Eren, O., Depew, B., & Barnes, S. (2017). Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics*, *153*(September), 9–31. https://doi.org/10.1016/j.jpubeco.2017.07.002

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal*, *44*(3), 594–629. https://doi.org/10.3102/0002831207306767

Garrett, R., & Hong, G. (2016). Impacts of grouping and time on the math learning of language minority kindergartners. *Educational Evaluation and Policy Analysis*, *38*(2), 222–244. https://doi.org/10.3102/0162373715611484

Hernandez, D. J. (2011). *Double Jeopardy: How third grade reading skills and poverty influence high school graduation* (The Annie E. Casey Foundation). https://www.aecf.org/m/resourcedoc/AECF-DoubleJeopardy-2012-Full.pdf

Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, *35*(5), 499–531. https://doi.org/10.3102/1076998609359785

Hong, G. (2012). Marginal mean weighting through stratification: A generalized model for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods*, *17*(1), 44–60. https://doi.org/10.1037/a0024918

Hong, G. (2015). *Causality in a social world: Moderation, mediation, and spill-over*. Wiley-Blackwell. https://doi.org/10.1002/9781119030638

Hong, G., Corter, C., Hong, Y., & Pelletier, J. (2012). Differential effects of literacy instruction time and homogeneous ability grouping in kindergarten classrooms: Who will benefit? Who will suffer? *Educational Evaluation and Policy Analysis*, *34*(1), 69–88. https://doi.org/10.3102/0162373711424206

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis*, *31*(1), 54–81. https://doi.org/10.3102/0162373708328259

Huddleston, A. P. (2014). Achievement at whose expense? A literature review of test-based grade retention policies in U.S. schools. *Education Policy Analysis Archives*, *22*(18). https://doi.org/10.14507/epaa.v22n18.2014

Jacob, B. A. (2003). A closer look at achievement gains under high stakes testing in Chicago. In P. E. Peterson, & M. R. West (Eds.), *No child left behind. The politics and practice of school accountability* (pp. 269–291). Brookings Institution Press.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, *89*(5–6), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, R. T., Stone, S., & Roderick, M. (2004). *Ending social promotion: Response of students and teachers*. Consortium on Chicago School Research. https://consortium.uchicago.edu/publications/ending-social-promotion-response-teachers-and-students

Johnson, S. M. (1986). Incentives for teachers: What motivates, what matters. *Educational Administration Quarterly*, *22*(3), 54–79. https://doi.org/10.1177/0013161X86022003003

Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)* (No. MR-792-PCT/FF). RAND. https://files.eric.ed.gov/fulltext/ED405365.pdf

Krieg, J. M. (2008). Are students left behind? The distributional effects of the No Child Left Behind. *Education Finance and Policy*, *3*(2), 250–281. https://doi.org/10.1162/edfp.2008.3.2.250

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effect s of school accountability policies. *Journal of Policy Analysis and Management*, *29*(3), 426–450. https://doi.org/10.1002/pam.20504

Mariano, L. T., Kirby, S. N., Crego, A., & Setodji, C. M. (2009). Measuring the effect of supportive interventions on proximal-year student achievement. In J. S. McCombs, S. N. Kirby, & L. T. Mariano (Eds.), *Ending social promotion without leaving children behind: The case of New York* (pp. 127–142). RAND.

Marsh, J. A., Gershwin, D., Kirby, S. N., & Xia, N. (2009). *Retaining students in grade: Lessons learned regarding policy design and implementation* (RAND Technical Report No. 677). https://doi.org/10.1037/e528822010-001

McCombs, J. S., Kirby, S. N., & Mariano, L. T. (Eds.). (2009). *Ending social promotion without leaving children behind: The case of New York*. RAND. https://doi.org/10.7249/RB9470

National Center for Education Statistics. (2002). *Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS–K), Psychometric report for kindergarten through first grade* (NCES 2002–05). Washington, DC. https://files.eric.ed.gov/fulltext/ED576531.pdf

National Center for Education Statistics. (2005). *Early Childhood Longitudinal Study—Kindergarten class of 1998-99 (ECLS-K): Psychometric report for the third grade* (NCES 2005-062). Washington, DC. https://nces.ed.gov/pubs2006/2006036rev.pdf

National Conference of State Legislatures. (2019). *Third-grade reading legislation*. http://www.ncsl.org/research/education/third-grade-reading-legislation.aspx

National Council of Teachers of Mathematics. (2009). *Focus in Grade 3: Teaching with curriculum focal points*.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00–4769). U.S. Government Printing Office. https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, *92*(2), 263–283. https://doi.org/10.1162/rest.2010.12318

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102/01623737026003237

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis*, *30*(2), 111–140. https://doi.org/10.3102/0162373708317680

Penfield, R. D. (2010). Test-based retention: Does it stands up to professional standards for fair and appropriate test use? *Educational Researcher*, *39*(2), 110–119. https://doi.org/10.3102/0013189X10363007

Polikoff, M. S. (2012). Instructional alignment under No Child Left Behind. *American Journal of Education*, *118*(3), 341–368. https://doi.org/10.1086/664773

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524. https://doi.org/10.1080/01621459.1984.10478078

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5–6), 1394–1415. https://doi.org/10.1016/j.jpubeco.2007.05.003

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving students to high-stakes testing. *Educational Evaluation and Policy Analysis*, *23*(3), 197–227. https://doi.org/10.3102/01623737023003197

Roderick, M., Jacob, B. A., & Bryk, A. S. (2002). The impact of high-stakes testing in Chicago on student achievement in promotional gate grades. *Educational Evaluation and Policy Analysis*, *24*(4), 333–357. https://doi.org/10.3102/01623737024004333

Roderick, M., & Nagaoka, J. (2005). Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis*, *27*(4), 309–340. https://doi.org/10.3102/01623737027004309

Rowan, B., Chiang, F., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, *70*(October), 256–284. https://doi.org/10.2307/2673267

Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*(August), 154–169. https://doi.org/10.1016/j.jpubeco.2017.06.004

Spillane, J. P. (2005). Primary school leadership practice: How the subject matters. *School Leadership & Management*, *25*(4), 383–397. https://doi.org/10.1080/13634230500197231

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics*

of Education Review, *27*(5), 556–563. https://doi.org/10.1016/j.econedurev.2007.06.004

Stipek, D. (2002). *Motivation to learn: Integrating theory and practice* (4th ed.). Allyn & Bacon.

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2001). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost*. CRESST/CREDE/University of Colorado at Boulder.

Thomas, M. R. (2005). *High-stakes testing: Coping with collateral damage*. Lawrence Erlbaum. https://doi.org/10.4324/9781410612809

Winters, M. A., & Greene, J. P. (2012). The medium-run effects of Florida's test-based promotion policy. *Education Finance and Policy*, *7*(3), 305–330. https://doi.org/10.1162/EDFP_a_00069

Wonder-McDowell, C., Reutzel, R. D., & Smith, J. A. (2011). Does instructional alignment matter? Effects on struggling second graders' reading achievement. *Elementary School Journal*, *112*(2), 259–279. https://doi.org/10.1086/661524

Xue, Y., & Meisels, S. J. (2004). Early literacy instruction and learning in kindergarten: Evidence from the Early Childhood Longitudinal Study—Kindergarten class of 1998-1999. *American Educational Research Journal*, *41*(1), 191–299. https://doi.org/10.3102/00028312041001191

Authors

YIHUA HONG is a researcher at the Center for Research, Evaluation and Equity in Education of RTI International. Her research interests include test-based accountability, new teacher support, and instructional effectiveness.

GUANGLEI HONG is a professor at the Department of Comparative Human Development and the College of the University of Chicago. She has focused her research on developing causal inference theories and methods for evaluating educational and social policies and programs in multilevel, longitudinal settings.