Learning from Adversity? Short- and Long-Term Spillover Effects from Grade Retention in Kindergarten☆

JOB MARKET PAPER

Jan Bietenbeck

CEMFI, Casado del Alisal 5, 28014 Madrid, Spain. Email: bietenbeck@cemfi.es

Abstract

Grade retention rates in kindergarten and the early elementary grades have risen steadily over the past few decades in the United States. While many studies document that retention impedes skill accumulation among retained students, little is known about the impact of retention policies on the outcomes of non-retained students. This study estimates the causal spillover effects from retained students on the cognitive and non-cognitive skills of their non-retained kindergarten peers. It draws on data from the Tennessee STAR experiment, which randomly assigned students to classes, and documents three sets of impacts. First, students exposed to retained classmates score lower on a standardized mathematics test at the end of kindergarten, an effect that fades out in later years. Second, exposed students score higher on a variety of measures of non-cognitive skills that are first observed about three years after kindergarten, and they seem to be able to retain these non-cognitive gains over time. Third, students benefit from kindergarten exposure to retained classmates in the long run, as they are more likely to graduate from high school and to take a college entrance exam. I argue that these favorable long-term effects are driven by greater non-cognitive skills such as improved discipline, which students acquired as they learned to cope with the initially adverse situation of being in class with an underachieving and potentially disruptive retained student.

1. Introduction

Grade retention rates in kindergarten and the early elementary grades have risen steadily over the past few decades in the United States (Hauser, Frederick, and Andrew, 2007). During the 2010-11 school year, for example, 6% of kindergarten

☆This version: December 2014. I am greatly indebted to David Dorn for his guidance and support. I thank David Autor, Manuel Arellano, Manuel Bagues, Sascha Becker, Samuel Bentolila, Bryan Graham, Pedro Mira, Magne Mogstad, Luca Repetto, Susan Dynarski, Jan Stuhler, Simon Wiederhold, and seminar audiences at CEMFI, the University of Zurich, the IZA Summer School, the SOLE and EALE conferences, the SAEe in Palma de Mallorca, and the European Winter Meeting of the Econometric Society for helpful comments. Diane Schanzenbach generously provided a subset of the data used in this paper. Funding from the Spanish Ministry of Science and Innovation (BES-2011-050947) is gratefully acknowledged.

students across the country were repeating the grade (NCES, 2013). This rise in the incidence of retention defies overwhelming empirical evidence showing that retained students obtain at best a transitory advantage, and more likely suffer considerable negative consequences in terms of cognitive and non-cognitive development, from being retained in grade.[1] While the effects of grade retention on retained students have been widely studied, little is known about how retention policies affect the learning outcomes of non-retained students.

In this paper, I analyze spillovers from grade-retained kindergarten students, henceforth called "repeaters," on their non-retained classmates. There are at least three channels through which repeaters can influence the learning outcomes of their non-retained peers. First, repeaters are typically academic underachievers, and a long line of research in economics of education has documented negative spillovers from low-achieving peers on their classmates (e.g. Burke and Sass, 2013). Second, a number of studies in educational psychology (e.g. Pagani et al., 2001) suggest that retention induces disruptive behavior in repeaters, and previous research in economics has documented negative spillovers from disruptive peers (e.g. Figlio, 2007). Finally, there may also be positive spillovers from repeaters to the extent that students learn from being in class with "failed" peers (Hill, 2014). As a particular case of this last channel, students may "learn from adversity" as the presence of a disturbing repeater requires them to become more persistent and disciplined in order to succeed in school.[2]

This paper provides empirical evidence on these potential spillovers using data from the Tennessee Student-Teacher Achievement Ratio (STAR) experiment, which randomly assigned kindergarten students and teachers to classes within schools. There are two features of these data that make them uniquely suited for the present research. First, any study of spillovers from repeaters faces the challenge that classes with and without repeaters might differ systematically in unobservable character-

[1] In an influential meta analysis on the topic, Jimerson (2001) concludes that research "fail[s] to demonstrate that grade retention provides greater benefits to students with academic or adjustment difficulties than does promotion to the next grade" (pp. 434-435). More recently, Hong and Raudenbush (2006) and Fruehwirth, Navarro, and Takahashi (2014) find substantial negative effects from kindergarten retention on the subsequent performance of retained students on standardized tests. In contrast, Dong (2010) reports short-term gains from kindergarten retention on standardized test scores, which however fade out over time.

[2] There are a variety of potential explanations for spillovers working through each of the three channels mentioned here. For example, spillovers through low academic achievement might be due to teachers having to slow down the speed of instruction because of the presence of underachieving repeaters. In contrast, an alternative to the "learning from adversity" explanation for positive spillovers is that being in class with a repeater induces non-repeating students to exert more effort in order not to be retained themselves.

istics. For example, if school principals tend to assign high-performing teachers to classes containing repeaters, any naive regression of student performance on class-level repeater exposure that does not control for this assignment process will lead to spillover estimates that are upward biased. The random assignment of students, including repeaters, and teachers to classes in Project STAR means that such biases are not present here. Second, students in the experiment were followed until the end of high school, which allows for the unique opportunity of studying long-term consequences of kindergarten repeater exposure.

The empirical analysis relates regular (non-repeating) students' kindergarten exposure to repeaters, measured as being randomly assigned to a class containing at least one repeating student, to their short- and long-term school performance. Repeater exposure reduces regular students' scores on an end-of-kindergarten math test by 9% of a standard deviation, which is a sizable effect for an educational intervention. A heterogeneity analysis reveals that this negative spillover effect is larger for African American students and children from poor families. Repeater exposure however substantially *increases* students' non-cognitive skills, which are first measured about three years after kindergarten. While the negative spillover effects from repeater exposure on test scores fade out after some years, the gains in non-cognitive skills persist over time. Strikingly, students who are exposed to repeating classmates in kindergarten appear to benefit from this experience in the long run, as they are more likely to graduate from high school and to take a college entrance exam.

To aid the interpretation of these results, I present additional evidence that allows to partially disentangle the mechanisms behind these spillover effects. I first show that it is likely the classroom disruption by repeaters, rather than their low academic ability, that is responsible for the negative impact of repeater exposure on regular students' kindergarten test scores. In contrast, a mechanism of learning from adversity could explain the positive spillover effects on long-term outcomes. In particular, one possible interpretation of the results is that exposed students learn to be persistent and disciplined in the face of adverse learning conditions due to classroom disruption by repeaters. These improved personality traits are reflected in higher non-cognitive skills. In line with abundant empirical evidence showing that non-cognitive skills acquired in early childhood are important for long-term educational outcomes (e.g. Heckman, Stixrud, and Urzua, 2006; Deming, 2009), the differential accumulation of such skills by exposed students then drives the positive long-term spillovers.

There is little previous research on spillover effects from repeating students. Lavy, Paserman, and Schlosser (2012) find that the share of old-for-grade students in a class negatively affects regular students' performance in Israeli high schools. Most of these old-for-grade students however entered school late, while only a minority are repeaters. The results of that paper thus relate to the literature on spillovers from low-achieving classmates more generally, rather than capturing specific effects of repeaters. Another paper by Hill (2014) finds that the share of repeaters in a student's high school math course positively affects her probability of failing that course. A key difference to the setting in this paper is the focus on high schools, where students typically repeat a specific course rather than an entire grade. Finally, Gottfried (2013) reports a negative effect of the number of repeaters in a class on non-repeating students' third- and fourth-grade test scores. A potential caveat of that study is that identification is based on the assumption that repeaters are randomly assigned to classes within schools conditional on observables.[3]

My contribution to this literature is threefold. First, random assignment of students and teachers to classes in Project STAR allows me to identify plausibly causal spillover effects, whereas the results of previous studies often do not permit a causal interpretation. Second, I estimate the impacts of repeaters on regular students' non-cognitive skills, which are receiving increased attention in research because of their important role for shaping long-term outcomes. Finally, whereas previous studies have focused exclusively on contemporaneous spillovers, I am able to move beyond the short-term perspective by tracing students' outcomes all the way from kindergarten to the end of high school. Thus, this paper provides the first results on long-term effects from kindergarten repeater exposure, and some of the first evidence on long-term effects from early childhood peers more generally.[4]

2. The STAR data

2.1. Background on Project STAR

Project STAR was a randomized experiment designed to study the effects of class size on student achievement. In the beginning of the 1985-86 school year, 6,325 kindergarten students in 79 participating Tennessee schools were randomly assigned to small (target size 13-17 students) or regular-sized (22-25 students) classes within their schools.[5] Students were supposed to stay in their assigned class type (small

[3]This paper is further related to a study by Babcock and Bedard (2011), who find that a one standard deviation increase in the state-of-birth retention rates in first and second grade is associated with a 0.7% increase in mean male hourly wages, and to the literature on peer effects in schools more generally (see Sacerdote (2011) for a recent overview).

[4]Gould, Lavy, and Paserman (2009) study the long-term effects from attending an elementary school class with a higher share of immigrants.

[5]There was also a third class type: regular-sized classes with a full-time teacher's aide. Like previous analyses of Project STAR, I do not find any differences between the impacts of regular-

versus regular-sized) until the end of third grade, after which the experiment ended and they would return to ordinary classes. Students that joined the initial cohort in participating schools after kindergarten were also randomly assigned to class types, as were teachers in each grade.

This study exploits the fact that kindergarten students, including repeaters, and teachers were randomly assigned not only to class type, but also to a particular class within each type (50 schools in the experiment had multiple classes per type). Early analyses of Project STAR were reluctant to conclude that this was indeed the case, mainly because the STAR Technical Report (Word et al., 1990) does not describe the exact procedure by which students were allocated to specific classes. However, several more recent studies (Chetty et al., 2011; Cascio and Schanzenbach, 2013; Sojourner, 2013) also rely on random assignment of students and teachers to classes within types in Project STAR and provide new evidence in support of this assumption. Section 3 revisits some of this evidence and provides additional statistical support for the claim that repeaters and non-repeaters were randomly assigned to kindergarten classes within schools.

The eventual implementation of Project STAR differed somewhat from the original experimental design. Three aspects are particularly noteworthy in this regard. First, as the initial cohort of students advanced from kindergarten to third grade, there was substantial attrition due to students moving to other schools or being retained in grade. Thus, by the time the cohort reached third grade, 49% of students treated in kindergarten had left the experiment. Second, because of complaints of some parents about their children's initial assignment, students in regular-sized classes were re-randomized at the beginning of first grade. Third, while compliance with treatment assignment was nearly perfect in kindergarten, approximately 10% of students managed to switch between small and regular-sized classes in each of the subsequent grades (Krueger, 1999).

Due to the focus on spillovers from repeaters in kindergarten, non-compliance with class assignment in the later grades does not affect the causal interpretation of results in this paper. In contrast, sample attrition could potentially confound some of the estimates on long-term outcomes. I test for selective attrition in Section 5 below, but do not find it to be a problem for the large majority of outcomes studied here. Finally, the three aspects of the implementation mentioned in the previous paragraph change the total amount of time that students spend in class with a kindergarten repeater. This affects the interpretation of the repeater exposure treatment, a point

sized classes with and without a full-time teacher's aide. I therefore follow the convention in the literature and group these two class types in the empirical analysis.

that I will discuss in more detail in the following subsection.[6]

2.2. Variable definitions

Data for students participating in Project STAR were collected by various research teams and organizations both during the experiment and in several rounds after the experiment ended. The Project STAR public use file, on which the empirical analysis below is based, combines these data such that students can be followed throughout their scholastic careers until the end of high school. This subsection gives a brief overview of the dependent and independent variables used in the empirical analysis. The data appendix provides additional details on data collection procedures and on the construction of outcome variables.

Demographic characteristics. The data contain information on students' gender, race, eligibility for free or reduced-price lunch, and exact date of birth. Children in Tennessee are supposed to enter kindergarten if they are five years old on or before September 30 of a given year, and I use this rule to construct an old-for-grade indicator that takes value 1 if the student was six years or older on September 30, 1985, and 0 otherwise. Kindergarten students in Project STAR can be old for grade for two reasons: either they entered school late - the so-called "red-shirting" - or they were repeating kindergarten.[7]

Kindergarten repeaters. The data include an indicator for whether each student was repeating kindergarten in the 1985-86 school year. There are 253 repeaters in the sample, 193 of whom are old for grade. Note that *all* repeaters would be expected to be old for grade if they had entered kindergarten in accordance with Tennessee's school entry rules during one of the previous school years. Therefore, the 60 repeaters who were not old for grade must have entered school early. The empirical analysis below focuses on spillover effects from the 193 old-for-grade repeaters, who first entered kindergarten at the regular entry age. While the data do not contain information on the reason why they had been retained in grade, these students had likely been identified by principals, teachers, or parents as having cognitive or behavioral deficiencies that would have put them at a disadvantage had they been promoted to first grade. The same is not necessarily true for the 60 other repeaters, who might have been retained only because they were too young to enter

[6]Additional details regarding the design and implementation of Project STAR can be found in Word et al. (1990), Krueger (1999), and Finn et al. (2007).

[7]See Deming and Dynarski (2008) for an analysis of the red-shirting phenomenon in the United States.

first grade.[8]

Figure 1 shows a histogram of the share of old-for-grade repeaters in kindergarten for each of the 79 schools in Project STAR. Nineteen schools contain no kindergarten repeater. Among the other schools, the mean share of repeaters is 4.2%, whereas the median share is 3.1%. There is a single outlier school in which more than 18% of kindergarten students are repeaters, a figure that is more than four standard deviations above the sample median for schools with positive repeater shares. A sensitivity analysis in Section 5 shows that results are robust to excluding this school from the sample. Compared to schools that do not contain repeaters, schools with positive repeater shares are slightly smaller (average enrollment of 73 students versus 83 students), are less likely to be located in the inner city (12% versus 47% of schools), and contain lower fractions of black students (20% versus 61%) and low-income students (41% versus 67%) on average.

Repeater exposure. Figure 2 shows the distribution of repeaters across classes in schools with positive repeater shares. 126 of the 254 classes in this subsample contain no repeater. Among the 128 other classes, about two thirds contain exactly one repeater, while there are few classes with two or more repeaters. In view of this heavily skewed distribution, the main specifications of the empirical analysis will distinguish just between classes with and without repeaters. As a robustness check, I also measure repeater exposure as the actual number of repeaters in class, or as the share of repeaters in class. Results from these alternative specifications suggest that outcomes are similar for students who are exposed to one or to several repeaters, which implies that the main specifications using a dummy variable for the presence of at least one repeater in class do not unduly miss heterogeneous treatment effects.

An important question for the interpretation of results is whether the spillovers on long-term outcomes documented in this paper arise from exposure to a repeater during kindergarten or from exposure over a longer time horizon. If all children had stayed in their assigned classes until the end of the experiment, regular students would have been exposed to kindergarten repeaters either for four years or not at all until third grade. In practice, however, due to the various deviations from

[8] Children are required to be six years old on September 30 of the year they start first grade. It seems reasonable to assume that this rule was enforced more strictly than the kindergarten entry rule since kindergarten attendance was not mandatory in Tennessee at the time of the experiment. Empirically, the 60 "young" repeaters come from more favorable demographic backgrounds and have better short- and long-term school outcomes than the 193 old-for-grade repeaters. If all 253 repeaters are included in the empirical analysis, the estimated spillover effects are somewhat smaller than the ones reported in the paper.

the original experimental design described above, students who were exposed to a repeater in kindergarten and who were traced through third grade ended up being in class with one of these repeaters for 2.4 years on average, whereas students not exposed to a repeater in kindergarten ended up being in class with a repeater for an average of 0.6 years.[9] The treatment studied in this paper thus consists of exposure to a repeating student in kindergarten and about one additional year of exposure to one of these repeaters during early elementary school.[10]

Short- and long-term outcomes. At the end of each grade level from kindergarten through third grade, students were administered the grade-appropriate version of the Stanford Achievement Test. In the spring of grades 5-8, all participants still attending public school in Tennessee took the Comprehensive Test of Basic Skills as part of a statewide student assessment program. Both tests are standardized multiple-choice assessments with components in mathematics and reading. The empirical analysis below presents estimates of the effects of kindergarten repeater exposure on test scores at four grade levels: (i) kindergarten, (ii) first grade, (iii) fifth grade, which corresponds to the final year of elementary school, and (iv) eighth grade, which corresponds to the final year of middle school.

In November 1989, when participants were in fourth grade, teachers in the STAR schools were asked to evaluate a random subset of their students on a set of behavioral measures. Teacher ratings were recorded on a scale from 1-5 and were consolidated into three indices. The effort index is based on such items as whether a student completes her homework and whether she is persistent when confronted with difficult problems. The initiative index captures such characteristics as whether a student participates actively in classroom discussions. Finally, the discipline index includes items such as whether a student often acts restless and whether she interferes with peers' work. In eighth grade, math and English teachers were asked to rate a different random subset of STAR participants on similar questions, the answers to which were consolidated into the same three indices. The total of six

[9]These figures are computed for the subsample of students in schools with positive repeater shares. Note that a complete history of repeater exposure cannot be determined for participants in Project STAR because students who left the experiment were no longer followed and because repeater status and previous repeater exposure were not recorded for students who entered the experiment after kindergarten.

[10]A potentially important concern is that students exposed to a repeater in kindergarten might have been "compensated" by being assigned to classes without repeaters in the later grades. This seems unlikely because the salient feature of Project STAR was the assignment of students to classes of different sizes and not to classes with and without repeaters. Nevertheless, I confirmed that students in classes with and without one of the original kindergarten repeaters in the sample are balanced on students' demographic characteristics at each grade level.

fourth- and eighth-grade indices derived from teacher ratings will serve as measures of non-cognitive skills in the empirical analysis below.

Most STAR participants graduated from high school in 1998, and transcripts including information on high school grade point average (GPA) and graduation status were collected from selected high schools in 1999 and 2000. Colleges and universities in the United States typically require applying students to report results from either the ACT or the SAT test. In 1998, Krueger and Whitmore (2001) matched all STAR students to the administrative records of the two companies responsible for these tests. The outcome of this process is an indicator that takes value 1 if a student took either of these college entrance exams in 1998 and 0 otherwise. This indicator can be interpreted as measuring students' intention to go to college.[11] Together, high school GPA, high school graduation status, and college test taking form the set of long-term outcomes studied in this paper.

2.3. Sample selection and descriptive statistics

The full sample includes 6,325 kindergarten students in 127 small and 198 regular-sized classes in 79 schools. I exclude 28 students for whom repeater status is not observed and five students with missing demographic characteristics from this sample. I further drop the 60 repeaters who are not old for grade. These students had likely been in class with one of the old-for-grade repeaters during the previous (1984-85) school year, which makes the interpretation of spillovers less straightforward for them. Finally, while schools without repeaters do not contribute to the identification of spillover effects in this paper, they are kept in the sample in order to increase the precision of the estimated impacts of other covariates included in the regressions. The final estimation sample thus consists of 6,232 students, 193 of whom are repeaters. Results in this paper are robust to relaxing the sample restrictions discussed in this paragraph.

Table 1 shows descriptive statistics for demographic characteristics, the repeater exposure measure, and short- and long-term outcome variables separately for non-repeating and repeating kindergarten students in the estimation sample. Note that students exhibit lower socioeconomic characteristics than the student populations in Tennessee and the United States as a whole because Project STAR oversampled

[11] Technically, a zero on this indicator means that students either did not take a college entrance exam or had been retained in grade and took the exam in later years. I estimated the effect of kindergarten repeater exposure on an indicator for having been retained by 1994, which is the last year that this outcome can be observed in the data. The estimates showed no effect of repeater exposure on subsequent grade retention, which favors an interpretation of the test-taking indicator as students' intention to go to college.

schools in low-income neighborhoods (Krueger and Whitmore, 2001). Repeaters in the sample are predominantly male and are more likely to be eligible for free or reduced-price lunch than non-repeating students. Repeaters are also older than non-repeating students by definition. Since low-income schools with primarily black student populations have lower repeater shares on average, the repeating students in the sample are less likely to be black. Finally, only three percent of non-repeating students are old for grade, which shows that red-shirting was not common in these schools at the time of the experiment.

In order to facilitate easy comparison between short- and long-term outcomes of regular students and repeaters, I standardize all test scores and non-cognitive skills indices to have mean 0 and standard deviation 1 across non-repeating students in the estimation sample. Table 1 shows that repeaters perform substantially worse in school than regular students both in kindergarten and later on. For instance, repeaters score almost half a standard deviation below non-repeating students on the end-of-kindergarten reading test. This test score gap widens to almost a full standard deviation by fifth grade and stays constant afterwards. Table 1 further shows that repeaters tend to have considerably lower non-cognitive skills and are less likely to graduate from high school and to take a college entrance exam. This suggests that at least in the context of Project STAR, repeaters do not tend to catch up with regular kindergarten students in terms of school performance.[12]

3. Identification strategy and validity of the experimental design

3.1. Identification based on between-class variation in repeater exposure

The random assignment of students to kindergarten classes in Project STAR led to a situation where within the same school, some classes contained repeaters while others did not. This between-class variation in repeater exposure forms the basis for identification of spillover effects from repeaters in this paper. A challenge arises because repeater exposure is positively correlated with class size. In particular, repeaters are more likely to be observed in regular-sized classes because (i) larger classes are more likely to contain at least one repeater when students are randomly assigned to classes, and (ii) the sample contains more regular-sized classes than small classes.[13] Previous analyses of Project STAR have documented large negative effects

[12]Repeaters are less likely to be observed than non-repeaters on some of the post-kindergarten outcomes, which may raise concerns about selective attrition driving this conclusion. Note, however, that repeaters are also much less likely to take a college entrance exam in 1998, an outcome that is observed for all students in the sample.

[13]Consider, for example, a school with the typical configuration of one small class of 15 students and two regular-sized classes of 23 students. If this school contains one repeater (the mode among

of class size on student outcomes (see Schanzenbach (2006) for an overview of these findings). Therefore, a regression of student performance on repeater exposure that does not control for class size will yield an estimate that is negatively biased. To address this concern, I control for class size in all of my regressions. I also conducted the entire empirical analysis separately for small and for regular-sized classes as a robustness check, and found similar though less precisely estimated results.

Section 4 presents estimates of the following empirical model:

$$y_{ics} = \alpha_s + \beta_1 \text{EXPOSURE}_{cs} + \beta_2 \text{CLASSSIZE}_{cs} + X_{ics}\gamma + \varepsilon_{ics}, \quad (1)$$

where y_{ics} is a short- or long-term outcome for non-repeating student i randomly assigned to kindergarten class c in school s, EXPOSURE_{cs} is an indicator for whether student i's class contains at least one repeater, CLASSSIZE_{cs} is the number of students in class c, and X_{ics} is a vector containing the student demographic characteristics shown in Table 1. Because random assignment to classes took place within schools, the model also controls for a vector of school fixed effects (α_s). Note that repeaters only act as treatment and are not treated themselves since this paper aims to estimate spillover effects from repeaters on non-repeating students. Additionally, by separating the initiators from the recipients of spillover effects, I avoid the mechanical bias in peer effects studies recently discussed by Angrist (2014). This bias arises if some students provide treatment for other students while at the same time being subject to treatment from these other students themselves. A robustness check below confirms that results in this paper are not driven by these mechanics.

3.2. Evidence on random assignment of repeaters

The key identification assumption underlying the model in equation 1 is that conditional on school fixed effects and class size, classes with and without repeaters do not differ systematically in any other way. In non-experimental data, this assumption could be violated if, for example, high-performing teachers are assigned to classes containing repeaters. In contrast, random assignment of students and teachers to classes in Project STAR ensures that this assumption holds here. Chetty et al. (2011) and Cascio and Schanzenbach (2013) provide evidence in support of random assignment by showing that classes are balanced on a wide range of student demographics and teacher characteristics. Here, I complement this evidence by evaluating whether repeaters were randomly assigned to classes within schools.

schools with positive repeater shares), this repeater has a 46/61 probability of being assigned to a regular-sized class and a 15/61 probability of being assigned to the small class.

As a first test for random assignment, I checked whether the within-school variation in repeater exposure observed in the data is consistent with a random allocation process. To that end, I performed a Monte Carlo simulation in which students were randomly assigned to classes within schools and in which the number and size of classes and the number of repeaters in each school were based on the actual data. I then computed the within-school standard deviation in repeater exposure as a measure of the variation used for identification in the re-randomized data. Across 1,000 replications, the median standard deviation was 0.381 with a narrow 90% empirical confidence interval of [0.369, 0.391]. This confidence interval comfortably contains the within-school standard deviation of 0.383 observed in the actual data.

As a second test for random assignment, I regressed an indicator taking value 1 if the student is a repeater and 0 otherwise on school and class fixed effects (omitting one class per school to avoid collinearity). Following the intuition described by Chetty et al. (2011), if assignment to classes was indeed random, then class indicators should not predict predetermined repeater status in this regression. Consistent with this idea, the p-value from an F test for the joint significance of the class fixed effects is 0.65, suggesting that repeater status is indeed balanced across classes.

Finally, I tested directly whether being exposed to a repeater predicts non-repeating students' demographic characteristics. Appendix Table 1 reports results from regressions of the five demographic characteristics available in the data on the repeater exposure dummy (panel A) and on the number of repeaters in class (panel B). All specifications also control for school fixed effects. Across the ten regressions, the estimated coefficients on the measures of repeater exposure are small and, with one exception, not statistically significant at conventional levels. Overall, the evidence presented here strongly suggests that repeaters were indeed randomly assigned to classes within schools in Project STAR.

4. The short- and long-term effects of kindergarten repeater exposure

4.1. End-of-kindergarten test scores

I begin the empirical analysis by examining the effects of repeater exposure on end-of-kindergarten test scores. To get a sense of the potential importance of spillovers from repeaters, Figures 3A and 3B show kernel density plots of math and reading scores separately for regular students with and without repeater exposure. In both plots, the distribution for students exposed to a repeater is shifted leftwards from the distribution for students without repeater exposure. This shift is somewhat larger in magnitude for math scores than for reading scores. Figures 3A and 3B thus suggest negative spillover effects from repeaters on regular students' kindergarten test scores. Note, however, that the evidence presented here could still be confounded by differences in student achievement levels between schools and between classes of different sizes within schools.

Table 2 presents regression estimates that control for these two potential confounders. Column 1 shows a significant negative effect of repeater exposure on math scores. The estimated coefficient implies that being in class with a repeater decreases regular students' math scores by 9.2% of a standard deviation on average. In comparison, being assigned to a small class (an average of 15 students) rather than a regular-sized class (23 students) is estimated to raise math scores by 15.2% of a standard deviation. Column 2 adds controls for students' demographic background. Due to the random assignment of students to classes, these controls do not change the coefficient estimate for the repeater exposure treatment but they slightly improve its precision. Finally, columns 3 and 4 show corresponding results for reading scores. In agreement with the visual evidence presented in Figures 3A and 3B, the estimated impact of repeater exposure is substantially smaller and not statistically significant in these specifications.

An important question is whether repeaters differentially affect students from different demographic backgrounds. To examine whether this is the case, Table 3 reports estimates from specifications for the end-of-kindergarten math score in which repeater exposure is interacted with the five demographic characteristics available in the data. The results in columns 1-4 point in the direction of larger impacts from repeaters on males, blacks, students on free lunch, and younger students, though none of these estimates is statistically significant at conventional levels. Note that males, blacks, and students who are eligible for free lunch also tend to do worse on average on standardized tests according to Table 2.

In order to further explore whether students from less favorable demographic backgrounds are particularly adversely affected by the presence of a repeater in their class, I combine the demographic variables into an index as follows. First, I define a composite test score as the average of the end-of-kindergarten math and reading scores. For each student, the composite scores of all other non-repeating students in the sample are then regressed on the five demographic characteristics in columns 1-4 of Table 3 (a total of 6,039 regressions). Finally, I standardize the predicted values from these regressions to have mean 0 and standard deviation 1. The outcome of this procedure is an index which weights demographic characteristics such that they predict students' performance on end-of-kindergarten tests.[14] As would be expected

[14]I use the 6,039 leave-me-out regressions rather than one regression for all students in order to avoid a mechanical correlation between the demographic background index and the outcome in

from the results in Table 2, this index is positively correlated with age and negatively correlated with being male, black, and eligible for free lunch.

Column 5 of Table 3 shows results from a regression of the end-of-kindergarten math score in which repeater exposure is interacted with the demographic background index. The main exposure effect, which by construction captures spillovers on students from an average demographic background, is very similar to the baseline estimate from Table 2 at -10.1%. The interaction term is positive and marginally significant ($p = 0.06$), suggesting that students from less favorable demographic backgrounds are more adversely affected by the presence of repeaters in their class. The coefficient of 5.7% implies that students with a demographic background index one standard deviation below the mean have 15.8% of a standard deviation lower math scores on average if exposed to a repeater. Overall, the results in this table therefore point towards substantial heterogeneity in the size of spillovers according to students' demographic background.[15]

4.2. Elementary and middle school test scores

Table 4 reports results from regressions of math and reading scores in later grades on kindergarten repeater exposure. Panel A presents estimates of the baseline specification in equation 1, while panel B presents estimates from specifications in which repeater exposure is interacted with the demographic background index. Columns 1 and 2 of panel A report small and insignificant positive effects of kindergarten repeater exposure on first grade math and reading scores. The lower bound of the 95% confidence interval around the estimated coefficient in the math regression is -5.0%, corresponding to about half the size of the effect of repeater exposure on the kindergarten math score. Columns 3-6 show that over time, the impacts from repeater exposure grow increasingly positive. Indeed, column 5 reports that being exposed to a repeater in kindergarten raises eighth-grade math scores by an average of 5.9% of a standard deviation, an effect that is marginally significant.

Panel B of Table 4 shows positive impacts of the interaction between kindergarten repeater exposure and the demographic background index on later test scores. These estimates suggest that students from less favorable demographic backgrounds, whose test performance in kindergarten was more adversely affected by the presence of a repeater, continue to perform worse than repeater-exposed students from favorable demographic backgrounds. For instance, students whose demographic background

regressions with the end-of-kindergarten math and reading scores as dependent variables.

[15]Results from regressions of the end-of-kindergarten reading score similarly point towards larger negative spillovers from repeaters on students from less favorable demographic backgrounds. These results are available upon request.

index is one standard deviation below the sample mean are estimated to have 5.4% of a standard deviation lower math scores in first grade if they were exposed to a repeater in kindergarten, an effect that later diminishes. In contrast, the equivalent effect for students with a one standard deviation above average index is already positive in first grade, and grows subsequently. Overall, the results in this table suggest that the negative spillover effects from kindergarten repeaters on test scores fade out over time and even become positive, particularly in the case of students from favorable demographic backgrounds.

4.3. Non-cognitive skills

Much of the literature on school organization focuses on cognitive test scores in order to evaluate the efficacy of educational interventions. A rapidly growing literature however emphasizes that non-cognitive skills are as important or even more important for students' success in later life, and that such skills are partly formed in school (e.g. Heckman, Stixrud, and Urzua, 2006; Chetty et al., 2011; Heckman, Pinto, and Savelyev, 2012). Table 5 reports estimates of the effects of kindergarten repeater exposure on non-cognitive skills. Columns 1-6 of panel A show positive spillovers from repeaters on regular students' effort, initiative, and discipline in both fourth grade, when these skills are first measured, and eighth grade. In each period, the impacts on effort - which, among other things, captures students' persistence - and discipline are particularly large.

Column 7 of panel A reports the effect of kindergarten repeater exposure on a summary index of non-cognitive skills. Following Kling, Liebman, and Katz (2007), this index is constructed by averaging the six standardized fourth- and eighth-grade indices for each non-repeating student and normalizing the resulting composite to have mean 0 and standard deviation 1.[16] Being exposed to a repeater in kindergarten raises non-cognitive skills, as measured by the summary index, by a highly significant 12.1% of a standard deviation. In comparison, being assigned to a small rather than to a regular-sized kindergarten class is estimated to increase the non-cognitive skills index by only 4.1% of a standard deviation, an effect that is not statistically significant (not shown in table).

Finally, panel B of Table 5 shows estimated coefficients on the interaction between repeater exposure and the demographic background index that are small compared to the main exposure effect and never statistically significant. Overall, being exposed to a repeater in kindergarten therefore seems to have strikingly different impacts

[16] If only fourth-grade or only eighth-grade non-cognitive skills are observed for a student, the average of the available skill variables is used.

on cognitive and non-cognitive skills at least in the short term. While the effect on cognitive skills, as measured by test scores, is initially negative but fades out relatively fast, the impact on non-cognitive skills is positive and large and persists at least until eighth grade, when these skills are last measured.

4.4. Long-term outcomes

The scholastic outcomes of STAR students were last tracked at the end of high school through collection of data on high school GPA, high school graduation, and taking of college entrance exams. Table 6 reports estimates from regressions of these long-term outcomes on kindergarten repeater exposure. Being in class with a repeater raises regular students' high school GPA by 0.7 points on a scale of 100 points (column 1 of panel A) and increases their likelihood to graduate from high school by 2.3 percentage points (column 2). Both of these effects are marginally significant. Column 3 shows that kindergarten repeater exposure significantly increases the likelihood of taking a college entrance exam by 3.3 percentage points, corresponding to a sizable 8% increase over the base rate of 41%. Finally, column 4 shows that kindergarten repeater exposure has a highly significant positive impact on a summary index of the three long-term outcomes, which is constructed in an equivalent way to the non-cognitive skills index in the previous subsection.[17]

Panel B shows results from specifications that interact kindergarten repeater exposure with the demographic background index. The main exposure effects are positive and quantitatively similar to the ones reported in panel A across the four specifications. Moreover, there is a positive impact of the interaction term on high school GPA, college test taking, and the long-term outcomes index, though the former is not statistically significant. The coefficients in column 4 imply that the effect of kindergarten repeater exposure on long term outcomes is close to zero for students with a demographic background index one standard deviation below the average, but positive and large for students from more favorable demographic backgrounds. Thus, being exposed to a repeater in kindergarten improves long-term outcomes for the majority of students in the sample, a result that stands in stark contrast to the negative contemporaneous effect on test scores found above.

[17] Each of the three long-term outcomes is first standardized by subtracting its mean and dividing by its standard deviation. In a second step, the average of these standardized outcomes is then normalized to have mean 0 and standard deviation 1 across all non-repeating students in the estimation sample. All available long-term outcomes are used for each student.

5. Robustness

5.1. Alternative measures of repeater exposure

The measure of repeater exposure used in the main analysis distinguishes between classes with and without repeaters, but does not further differentiate repeater classes according to the number of repeaters. This measure is motivated by the fact that relatively few classes in the sample contain more than one repeater. Nevertheless, I now examine whether spillovers are larger in classes with more repeaters.

Panel A of Appendix Table 2 reports estimates from specifications that include separate indicators for being in class with 1, 2, and 3-5 repeaters. Estimates in this table are presented for four main outcomes: kindergarten math score, eighth-grade math score, the non-cognitive skills index, and the long-term outcomes index. Across all specifications, the estimated impacts from exposure to one and exposure to two repeaters are qualitatively and quantitatively similar to the headline results in Section 4. In contrast, the coefficients on being in class with three to five repeaters are smaller in absolute value but very imprecisely estimated due to the small number of classes containing more than two repeaters. One cannot reject the null hypothesis that spillovers are of equal size in all classes with repeaters, irrespective of the number of repeaters in class.

Panel B of Appendix Table 2 shows results from specifications which use the class share of repeaters as treatment and which are qualitatively similar to the results reported in Section 4. Finally, panel C reports results from regressions in which repeater exposure is measured by indicators for class shares of repeaters between 0.1-5%, 5-10%, and more than 10%. The estimates from these specifications mirror those in Panel A: the impacts from classes containing 0.1-5% and 5-10% of repeaters are overall similar to those reported in Section 4. While the results for being in class with more than 10% of repeaters are also qualitatively similar, they are mostly imprecisely estimated due to the small number of classes containing such high shares of repeaters. Overall, the findings reported in this table offer further justification for measuring repeater exposure using a dummy variable, as done in the main analysis.

5.2. Sample selection

Appendix Table 3 reports estimates from regressions that probe the robustness of results to a variety of sample restrictions. Panel A shows that similar though less precise estimates are obtained for the four main outcomes if the sample is restricted to regular-sized classes. This confirms that the potential bias due to the correlation between class size and repeater exposure is avoided by controlling for class size in the main specifications in Section 4. The similarity of results is also interesting in its own right because it suggests that spillovers from repeaters are not moderated by class size. To the extent that smaller classes allow teachers to better respond to the individual needs of each student, one might have hypothesized that the negative impact of repeater exposure on kindergarten math scores is attenuated in these classes.[18] The results reported here show that at least in the setting of Project STAR, this is not the case.[19]

Panel B presents results from specifications in which the sample is restricted to schools that contain at least one repeater. While schools without repeaters do not contribute to the identification of spillover effects in this paper, they were kept in the estimation sample in order to increase the precision of the estimated coefficients on control variables. As the estimates in panel B show, results are robust to dropping these schools from the sample. Finally, panel C reports estimates from a sample that excludes the outlier school in which more than 18% of students were repeating kindergarten (see Figure 1). As can be seen from the table, keeping this school in the sample hardly affects the results in this paper.

5.3. Selective attrition based on repeater exposure

One challenge for the analysis of long-term outcomes is that estimates could be compromised if students differentially attrit from the sample based on assignment to a kindergarten class with or without repeaters. For example, to the extent that repeater-exposed students from less favorable demographic backgrounds are less likely to be observed with long-term outcomes, this could explain some of the positive long-term spillover effects documented in Section 4. I now provide evidence that this mechanism is unlikely to drive the results in this paper.

One simple test for selective attrition is to estimate the effect of repeater exposure on the likelihood that a given outcome variable is observed for a student.[20] Intuitively, if there are important differences in follow-up rates between exposed and

[18] This hypothesis would be in line with the theoretical predictions by Lazear (2001), for example.

[19] I also estimated specifications for the unrestricted estimation sample that included an interaction between repeater exposure and class size. The estimated coefficients on the interaction term were mostly small and never statistically significant in these regressions, confirming that class size does not moderate repeater spillovers. Moreover, I confirmed that qualitatively similar results are obtained if the sample is restricted to small classes. However, due to the low number of schools exhibiting variation in repeater exposure between small classes (19 schools), coefficients in these specifications were less precisely estimated.

[20] There are three broad reasons why a student may not observed with a given outcome. First, students who were retained in grade have missing values for subsequent test scores in grades 1-3 as well as for non-cognitive skills. Second, students who moved away from participating schools during the experiment or who left the Tennessee public school system altogether are not observed with at least some of the later test scores and non-cognitive skills. Third, students might simply not have been selected for data collection for a given outcome.

non-exposed students, selection bias due to differential attrition might be a serious concern. Panel A of Appendix Table 4 presents estimates of the model in equation 1 in which the dependent variables are indicators for being observed with each short- and long-term outcome.[21] The estimated follow-up differentials are small and not significantly different from zero in all but one specification. Moreover, there is no clear pattern regarding whether exposed students are systematically more or less likely to be observed with these outcomes. Finally, note that the table does not present results for the college test taking indicator because it is never missing by construction, which implies that the positive effect of repeater exposure on this outcome cannot be explained by selective attrition.

While the overall attrition differentials between exposed and non-exposed students are very small, it might still be the case that the composition of students between the two groups changed over time. If, for example, exposed students from less favorable demographic backgrounds were less likely to be observed than non-exposed students from similar demographic backgrounds with a given outcome, this could explain a positive effect of repeater exposure on this outcome. One way to test for these kinds of patterns is to re-estimate the specifications in panel A while additionally including an interaction between repeater exposure and the demographic background index. Panel B presents results from these regressions. The estimated coefficients on both the main exposure effect and the interaction term are small and not significantly different from zero in all but two cases, suggesting that there are few systematic differences between exposed and non-exposed students observed with short- and long-term outcome variables.

Finally, another simple test for selective attrition is to check whether the main results of this paper can be replicated using only those students who are observed with an outcome that suffers from attrition. Panel C presents results from the corresponding regressions of the kindergarten math score, while panel D presents results from regressions of the ACT/SAT test taking indicator. Across the sixteen specifications, the vast majority of estimates are qualitatively and quantitatively similar to the corresponding main result in Section 4, even though some of them are less precisely estimated due to smaller sample sizes. Overall, the results in this table do not support the notion that selective attrition based on kindergarten repeater exposure biases the results in this paper.[22]

[21]For conciseness, Appendix Table 4 presents results for being observed with math scores only. Given that the difference between the numbers of students taking math and reading tests at each grade level are very small (see Table 1), results for reading scores are very similar.

[22]Following Lee (2009), I also estimated treatment effect bounds that are robust to selective attrition. Computing these bounds involves trimming the sample such that follow-up rates for

5.4. Measurement of non-cognitive skills relative to repeaters

Section 4 reports positive impacts from kindergarten exposure to repeaters on non-cognitive skills in fourth and eighth grade. A potential concern is that these improvements might simply reflect higher teacher ratings of regular students' behavior *relative* to the behavior of repeaters in the same class. I address this issue by re-estimating the impacts of repeater exposure on fourth-grade non-cognitive skills for the subset of students whose fourth-grade classes did not contain any of the 193 initial kindergarten repeaters. As Appendix Table 5 shows, the effects of repeater exposure in these regressions are slightly attenuated compared to those reported in Table 5 but qualitatively similar. The data do not allow me to observe classroom composition during eighth grade, which makes a similar robustness exercise for eighth-grade non-cognitive skills unfeasible. It should however be noted that by that time, the majority of students had switched to different (middle) schools, and it is not obvious that previously exposed students would be systematically paired with repeaters at these new schools. Overall, the arguments made here therefore do not support the idea that the positive impacts of repeater exposure on non-cognitive skills capture only a mechanical effect due to relative teacher ratings.

5.5. Testing for mechanical spillover effects

In a recent paper, Angrist (2014) documents a mechanical bias in peer effects studies which arises if some students provide treatment for other students while being subject to treatment from these other students themselves. Intuitively, that bias should be avoided in this paper due to the separation of initiators and recipients of spillover effects. I confirm this intuition in a simulation-based falsification test similar to the one developed by Feld and Zoelitz (2014). In particular, I exchange each student's classmates with a new set of peers randomly drawn from other classes in the same school. In this way, all students are assigned to a group of placebo classmates with whom they did not interact in their real-world classroom. I then re-estimate the effect of repeater exposure, measured using the placebo classmates, on end-of-kindergarten math scores. Any effect of repeater exposure in this regression reflects purely mechanical forces. In 1,000 replications of this exercise, the median coefficient on repeater exposure was -0.007 with a 90% empirical confidence interval of [-0.039, 0.030], which excludes the coefficient of -0.093 found in the actual data.

exposed and non-exposed students are equal. The trimming proportions are the differentials estimated in panel A of Appendix Table 4. Given the small differentials reported there, it is not surprising that the estimated bounds were generally very similar to the headline effects reported in Section 4.

This confirms that the mechanical forces described by Angrist (2014) do not bias the results in this paper.

6. Discussion and Mechanisms

The results reported in Section 4 indicate important spillovers from kindergarten repeaters on the outcomes of their non-repeating classmates. Initially, students exposed to a repeater score worse on end-of-kindergarten tests, an effect that is particularly pronounced for students from less favorable demographic backgrounds. However, this negative impact fades out relatively fast and by the end of eighth grade, the effect of kindergarten repeater exposure on test scores has turned positive for the majority of students. The impact of repeater exposure on non-cognitive skills is positive both when these are first measured about three years after kindergarten and later on at the end of middle school. Students who were exposed to a repeater appear to make particularly strong gains on non-cognitive skills related to effort and discipline. Consistent with these results, there are lasting positive long-term effects, as repeater-exposed students obtain better high school grades, are more likely to graduate from high school, and more likely to take a college entrance exam. I discuss possible interpretations for these results below.

6.1. Spillovers through low academic ability or through classroom disruption?

The introduction suggested two possible channels for negative spillovers from repeaters on their classmates' test scores. First, because repeaters are academic underachievers, they may force teachers to slow down the pace of instruction, thus creating a negative spillover effect on other students that operates through repeaters' inferior school performance. Second, repeaters may be more likely to display externalizing behavior that disrupts the classroom and distracts other students. One exercise that may help distinguish between these two mechanisms is to estimate separately the spillovers from male and female repeaters. There is an extensive literature documenting large gender gaps in disruptive behavior between boys and girls. For example, Bertrand and Pan (2013) report that already at the beginning of kindergarten, boys score 44% of a standard deviation higher than girls on a teacher rating of externalizing behavior. Therefore, to the extent that the negative spillovers on kindergarten test scores found above are driven by male repeaters, this might suggest that classroom disruption is the underlying mechanism.

Appendix Table 6 reports estimates from regressions of the end-of-kindergarten math and reading scores on separate indicators for being in class with a male and being in class with a female repeater. In these regressions, the sample is restricted to classes that contain either no repeater or only male or only female repeaters. Column 1 shows a negative and highly statistically significant effect of male repeater exposure on kindergarten math scores that is almost twice as large as the overall exposure effect found in Table 2. In contrast, the impact of female repeater exposure is less than a fifth of that in size and not statistically significant. Column 2 shows corresponding results for reading that point to similar heterogeneity by repeaters' gender. While the coefficient estimates from these regressions strongly suggest that male repeaters are responsible for the negative spillover effects found in the overall analysis, the gender-specific estimates are not sufficiently precise to reject the null hypothesis of equal impacts from male and female repeaters.

If one takes this evidence to suggest that male repeaters are driving the negative spillovers on kindergarten test scores, this result is only informative about the underlying mechanism to the extent that male repeaters are not at the same time lower academic achievers than female repeaters. As the lower part of Appendix Table 6 reports, it turns out that male repeaters score substantially *higher* than female repeaters on the end-of-kindergarten math and reading tests, though the latter difference is not statistically significant. The table further shows that male repeaters receive much lower teacher ratings on the fourth-grade discipline index, which among other things measures classroom disruption. These findings, together with the result that male repeaters appear to drive the negative spillover effects on kindergarten test scores, are consistent with the notion that classroom disruption is likely the main mechanism for these impacts.

6.2. Learning from adversity, non-cognitive skills, and long-term outcomes

An important question is how the negative spillovers on kindergarten test scores can be reconciled with the positive spillovers on later test scores, on non-cognitive skills, and on long-term outcomes. The introduction suggested a third possible channel through which regular students could be affected by repeaters, namely a positive spillover effect through learning from adversity. More specifically, students exposed to repeaters in kindergarten may learn to be persistent and disciplined in the face of adverse learning conditions due to exposure to a disruptive repeater. These personality traits could notably be reflected in the higher effort and discipline scores found in Table 5. In turn, greater accumulation of such non-cognitive skills may eventually help students improve test scores, and may drive the differentially better long-term outcomes of students who were exposed to repeaters in kindergarten. This explanation is consistent with abundant evidence on the importance of non-cognitive skills acquired in early childhood for long-term educational outcomes (e.g. Heckman, Stixrud, and Urzua, 2006; Deming, 2009).

To the extent that exposed students' higher persistence and discipline makes them resilient to disturbing repeaters, one implication of the learning from adversity interpretation is that these students should be less negatively affected than non-exposed students by the presence of a repeater in later grades. The re-randomization of students in regular-sized classes after kindergarten potentially provides an opportunity to examine this issue empirically. In particular, the re-randomization generated first-grade classrooms comprising both students with and without repeater exposure in kindergarten. Moreover, randomization implied that some of these newly composed classes contained a kindergarten repeater while others did not. Unfortunately, however, repeater exposure in first grade can only be measured with substantial error due to the arrival of a large number of new students for whom repeater status is not observed (see Section 2).[23]

I nevertheless estimated the effect of exposure to one of the original kindergarten repeaters during first grade on the first-grade math scores of students in regular-sized classes. The regression allowed the impact from repeaters to vary between previously exposed students and students without repeater experience by including an interaction term between exposure in kindergarten and exposure in first grade. The results from this exercise were consistent with the interpretation of learning from adversity: while students without previous repeater exposure were negatively affected by being in class with a repeater, the impact was positive for students who had been in class with a repeater in kindergarten. Due to the limitations of this analysis mentioned in the previous paragraph, however, this evidence should be taken as suggestive, and not conclusive, of a learning from adversity mechanism.

The interpretation suggested here hypothesizes that the positive spillovers from repeaters on long-term outcomes operate through improvements in students' non-cognitive skills. A statistical implication of this argument is that in a regression analysis of long-term outcomes, the coefficient on repeater exposure should be substantially attenuated when non-cognitive skills are added to the regression as (endogenous) control. Table 7 presents evidence consistent with this idea. Column 1 shows that the non-cognitive skills index is highly predictive of the index of long-term outcomes. Column 2 indicates that this correlation hardly changes if cognitive skills, as measured by the kindergarten composite score, are controlled for in the regression. Thus, the non-cognitive skills index appears to pick up abilities that are not fully reflected in early-life standardized test scores.

[23] A total of 2,314 students entered the participating schools in first grade. This compares to 4,495 students who participated in Project STAR during kindergarten and are still observed in first grade.

Column 3 reports an effect of kindergarten repeater exposure on long-term outcomes for the subsample of students observed with non-cognitive skills that is very similar to the one reported for the unrestricted estimation sample in Table 6. Column 4 shows that this effect is substantially attenuated and no longer statistically significant once the non-cognitive skills index is included as an additional regressor. Indeed, a Wald test rejects the null of equal coefficients on repeater exposure across the specifications in columns 3 and 4 with $p<0.01$. Column 5 indicates that controlling for the kindergarten composite score rather than the non-cognitive skills index does not lead to a similar reduction in the coefficient on repeater exposure. This suggests that unlike non-cognitive skills, early-life cognitive skills are not an important channel for repeater spillovers on long-term outcomes. Finally, as column 6 shows, controlling for both cognitive and non-cognitive skills once again leads to a substantial and statistically significant reduction in the repeater exposure coefficient.

Overall, the results in Table 7 are consistent with the idea that kindergarten repeater exposure improves long-term outcomes due to its positive impact on non-cognitive skills via learning from adversity. Importantly, even though this interpretation is supported by several pieces of empirical evidence presented in this section, I cannot firmly exclude that repeater exposure improves long-term outcomes through a different mechanism that is also correlated with gains in non-cognitive skills.

7. Conclusion

Grade retention policies have become increasingly popular in the United States during recent decades, but little is known about how these policies affect the large majority of students who are never retained, but who share a classroom with the retained students. This paper estimates spillover effects from kindergarten repeaters on regular students' short- and long-term school outcomes. It substantially improves on the scarce previous evidence on repeater spillovers, and on the evidence on peer effects more generally, by exploiting various attractive data features of Project STAR. In this experiment, students and teachers were randomly assigned to classes within schools, which facilitates the identification of causal spillover effects. Moreover, the long-term tracking of students and the availability of both cognitive and non-cognitive outcomes allow for a much richer analysis than just the study of contemporaneous test scores, which is typical in the peer effects literature.

The results indicate that regular students who are exposed to a repeater in their kindergarten class perform worse on end-of-kindergarten math and reading tests. In contrast, these students display substantially higher non-cognitive skills when such skills are first measured about three years after kindergarten. While the negative spillover effects from kindergarten repeaters on cognitive test scores fade out after some years, the gains in non-cognitive skills appear to persist. The favorable development of students who were exposed to a repeater in kindergarten culminates in significantly raised propensities to graduate from high school and to take a college entrance exam around the age of eighteen. These spillovers from kindergarten repeaters differ considerably across subgroups of students, with students from more favorable demographic backgrounds being less negatively affected by repeaters in the short term and benefiting more in the long term.

One possible interpretation of these results is that students learn from adversity, i.e. they become persistent and disciplined in the face of adverse learning conditions due to classroom disruption by repeaters. I provide suggestive evidence in favor of that interpretation: the initial negative spillovers on test scores appear to work through classroom disruption by repeaters, while the increase in non-cognitive skills, which include measures of persistence and discipline, appears crucial to explain the positive impact of early repeater exposure on long-term outcomes.

While the results in this paper suggest that overcoming early-life adversity can help children succeed in school in the long run, it is important to note that the dosage of the treatment studied here is relatively low: students typically were in class with a single repeater in kindergarten and in about one additional year during elementary school. Whether similar long-term benefits arise if students are exposed to a large number of repeaters in their class is a question for future research.

Data appendix

The Tennessee State Department of Education entrusted a consortium of researchers from four Tennessee universities and various state institutions with the planning and implementation of Project STAR. Even after the experiment ended, some members of this consortium - often with the help of other researchers - continued to collect data on outcomes of participating students for their own research. Finn et al. (2007) provide a detailed account of these data collection efforts. The public use file, on which the empirical analysis in this paper is based, combines these data such that students can be followed throughout their scholastic careers until the end of high school. Additional data on test scores in fifth and eighth grade were generously provided to me by Diane Schanzenbach. In what follows, I discuss in detail how the outcome variables used in the empirical analysis are constructed.

Test scores. The kindergarten and first-grade test scores used in this paper come from the grade-appropriate version of the Stanford Achievement Test, which was administered to participating students at the end of the respective school years. Fifth- and eighth-grade test scores come from the Comprehensive Test of Basic Skills, which was administered to students as part of a statewide testing program, and were collected from administrative records of the Tennessee State Department of Education. The public use file contains math and reading scores for students who attended fifth and eighth grade in 1991 and 1994, respectively. Therefore, test scores are not observed for a substantial number of students who had been retained in grade by those years. In contrast, the data supplied by Diane Schanzenbach contains students' test scores in 1991 and 1994 irrespective of the grade attended. Since scale scores in this data are comparable across grade levels (Finn et al., 2007), I focus on these latter two sets of test scores in the main analysis. Results are however robust to using only the test scores available in the public use file. I standardize all elementary and middle school test scores to have mean 0 and standard deviation 1 across all non-repeating kindergarten students in the estimation sample.

Non-cognitive skills. A first set of non-cognitive skill measures comes from a questionnaire administered to teachers of a random sample of fourth-grade students in participating schools in November 1989. The questionnaire asked teachers to rate how often each student had engaged in 31 different behaviors over the last two to three months. Ratings were recorded on a scale from 1 ("never") to 5 ("always"). The answers to 28 of these behaviors were consolidated into four indices measuring each student's effort, initiative, discipline, and how much she valued school. The main analysis focuses on the first three indices, for which corresponding eighth-grade indices are available.

The effort index includes items such as whether a student is persistent when confronted with difficult problems, whether she completes her homework, and whether she gets discouraged easily when encountering an obstacle in schoolwork. The discipline index is based on such items as whether a student participates actively in classroom discussions, whether she does more than just the assigned work, and whether she often asks questions. Finally, the discipline index captures such characteristics as whether a student often acts restless, whether she needs reprimanding, and whether she interferes with peers' work.[24]

During the 1993-94 school year, eighth-grade math and English teachers of a different random subset of participants were asked about student behaviors on a similar though shorter questionnaire. Eleven of these behaviors were again consoli-

[24] See Finn et al. (2007) for a complete listing of the behaviors included in each of the indices.

dated into three indices measuring each student's effort, initiative, and discipline. I first average these three indices across math and English teachers for each student, and then normalize each of the effort, initiative, and discipline indices in fourth and eighth grade by subtracting its mean and dividing by its standard deviation. Finally, I construct the non-cognitive skills index by averaging these normalized indices and standardizing the resulting composite to have mean 0 and standard deviation 1 across all non-repeating students in the estimation sample.

High school GPA and graduation status. Most students in Project STAR graduated from high school in 1998, and transcripts were gathered from selected high schools in 1999 and 2000. High schools were chosen for data collection based on the likelihood that STAR participants would attend them given the locations of students' last known middle schools. Course grades from transcripts were transferred to a scale from 0-100 if necessary, and separate GPAs for math, science, and foreign language were computed and are available in the data. The empirical analysis in this paper uses the overall GPA, defined as the average of the these three subject-specific GPAs, as an outcome variable.

Information on high school graduation status was also derived from transcripts and verified with the Tennessee State Department of Education in ambiguous cases. Nevertheless, graduation status could not be determined with certainty for all students. In these cases, which comprise 7% of the non-repeating students in the estimation sample, the data collectors made a best guess whether a student "probably graduated" or "probably dropped out" based on the available course grades, information on attendance, and additional information from the Tennessee State Department of Education. The variable used in the empirical analysis codes 2,378 students who graduated, 98 students who probably graduated, and 82 students who received a General Educational Development certificate as graduates, and 296 students who dropped out and 101 students who probably dropped out as dropouts.

College test taking and long-term outcomes index. ACT/SAT test taking was recorded by Krueger and Whitmore (2001), who matched all students in STAR to the administrative records of the two companies responsible for these tests in 1998. The outcome variables used in the empirical analysis is an indicator that takes value 1 if a student took either of these college entrance exams in 1998 and 0 otherwise. The long-term outcomes index, which combines high school GPA and graduation and ACT/SAT test taking, is constructed by first normalizing each of these variables by subtracting its mean and dividing by its standard deviation. The average of these normalized outcomes is then standardized to have mean 0 and standard deviation 1 across all non-repeating students in the estimation sample.

References

Angrist, J.D. 2014. "The Perils of Peer Effects." *Labour Economics* 30:98–108.

Babcock, P., and K. Bedard. 2011. "The Wages of Failure: New Evidence on School Retention and Long-Run Outcomes." *Education Finance and Policy* 6:293–322.

Bertrand, M., and J. Pan. 2013. "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior." *American Economic Journal: Applied Economics* 5:32–64.

Burke, M.A., and T.R. Sass. 2013. "Classroom Peer Effects and Student Achievement." *Journal of Labor Economics* 31:51–82.

Cascio, E.U., and D.W. Schanzenbach. 2013. "First in the class? Age and the Education Production Function." Unpublished.

Chetty, R., J.N. Friedman, N. Hilger, E. Saez, D.W. Schanzenbach, and D. Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *The Quarterly Journal of Economics* 126:1593–1660.

Deming, D. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1:111–134.

Deming, D., and S. Dynarski. 2008. "The Lengthening of Childhood." *Journal of Economic Perspectives* 22(3):71–92.

Dong, Y. 2010. "Kept Back to Get Ahead? Kindergarten Retention and Academic Performance." *European Economic Review* 54:219–236.

Feld, J., and U. Zoelitz. 2014. "Understanding Peer Effects: On the Nature, Estimation and Channels of Peer Effects." Unpublished.

Figlio, D.N. 2007. "Boys Named Sue: Disruptive Children and Their Peers." *Education Finance and Policy* 2:376–394.

Finn, J.D., J. Boyd-Zaharias, R.M. Fish, and S.B. Gerber. 2007. "Project STAR and Beyond: Database User's Guide." Report, HEROS Incorporated.

Fruehwirth, J.C., S. Navarro, and Y. Takahashi. 2014. "How The Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects." Unpublished.

Gottfried, M.A. 2013. "The Spillover Effects of Grade-Retained Classmates: Evidence from Urban Elementary Schools." *American Journal of Education* 119:1–64.

Gould, E.D., V. Lavy, and M.D. Paserman. 2009. "Does Immigration Affect the Long-Term Educational Outcomes of Natives? Quasi-Experimental Evidence." *The Economic Journal* 119:1243–1269.

Hauser, R.M., C.B. Frederick, and M. Andrew. 2007. "Grade Retention in the Age of Standards-Based Reform." In A. Gamoran, ed. *Standards-based Reform and the Poverty Gap: Lessons for No Child Left Behind*. Brookings Institution Press, pp. 120–153.

Heckman, J., R. Pinto, and P. Savelyev. 2012. "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes." *American Economic Review* 103:2052–2086.

Heckman, J., J. Stixrud, and S. Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24:411–482.

Hill, A.J. 2014. "The Costs of Failure: Negative Externalities in High School Course Repetition." Unpublished.

Hong, G., and S.W. Raudenbush. 2006. "Evaluating Kindergarten Retention Policy." *Journal of the American Statistical Association* 101:901–910.

Jimerson, S.R. 2001. "Meta-Analysis of Grade Retention Research: Implications for Practice in the 21st Century." *School Psychology Review* 30:420–437.

Kling, J.R., J.B. Liebman, and L.F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75:83–119.

Krueger, A.B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114:497–532.

Krueger, A.B., and D.M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal* 111:1–28.

Lavy, V., M.D. Paserman, and A. Schlosser. 2012. "Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom." *The Economic Journal* 122:208–237.

Lazear, E.P. 2001. "Educational Production." *The Quarterly Journal of Economics* 116:777–803.

Lee, D.S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies* 76:1071–1102.

NCES. 2013. "Digest of Education Statistics 2012." Report, National Center for Education Statistics.

Pagani, L., R.E. Tremblay, F. Vitaro, B. Boulerice, and P. McDuff. 2001. "Effects of Grade Retention on Academic Performance and Behavioral Development." *Development and Psychopathology* 13:297–315.

Sacerdote, B. 2011. “Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?” Elsevier, vol. 3 of *Handbook of the Economics of Education*, pp. 249 – 277.

Schanzenbach, D.W. 2006. “What have researchers learned from Project STAR?” *Brookings Papers on Education Policy* 9:205–228.

Sojourner, A. 2013. “Identification of Peer Effects with Missing Peer Data: Evidence from Project STAR.” *The Economic Journal* 123:574–605.

Word, E., J. Johnston, H.P. Bain, D.B. Fulton, C.M. Achilles, M.N. Lintz, J. Folger, and C. Breda. 1990. “The State of Tennessee’s Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985-1990.” Report, Tennessee State Department of Education.

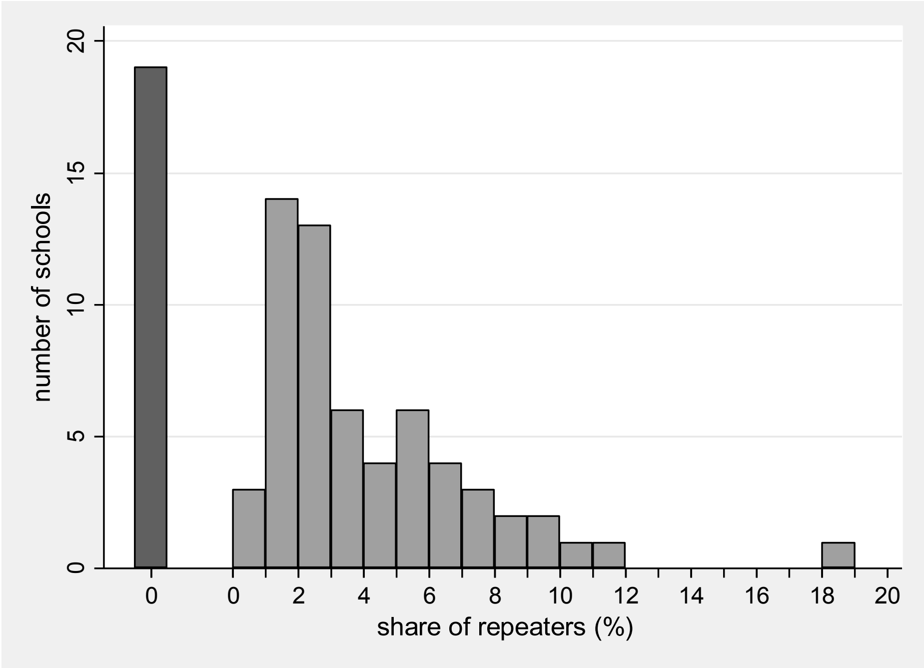

Figure 1: Distribution of repeaters across schools

Notes: The figure displays a histogram of the share of repeaters in kindergarten for each of the 79 schools in the data. The leftmost bar shows that there are 19 schools in which no student is repeating kindergarten. The remaining bars cover intervals of width 1%, starting with the interval]0, 1]. The mean (median) share of repeaters across schools with positive shares of repeaters is 4.2% (3.1%).

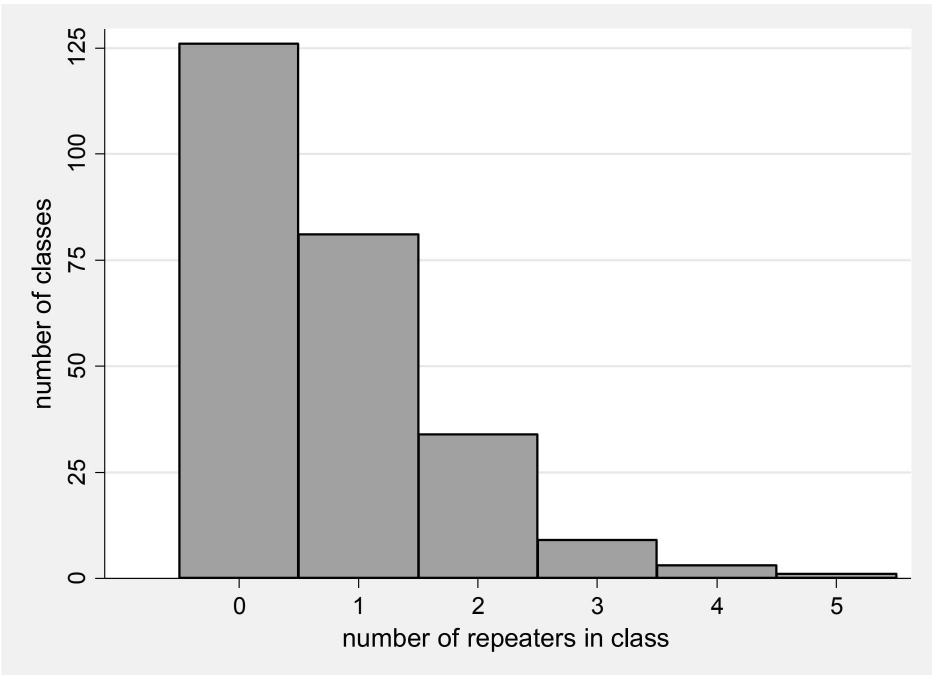

Figure 2: Distribution of repeaters across classes

Notes: The figure displays a histogram of the number of repeaters in class. The sample only includes schools with at least one repeater in kindergarten. There are 254 classes in this sample. The mean (median) number of repeaters in class is 0.76 (1).

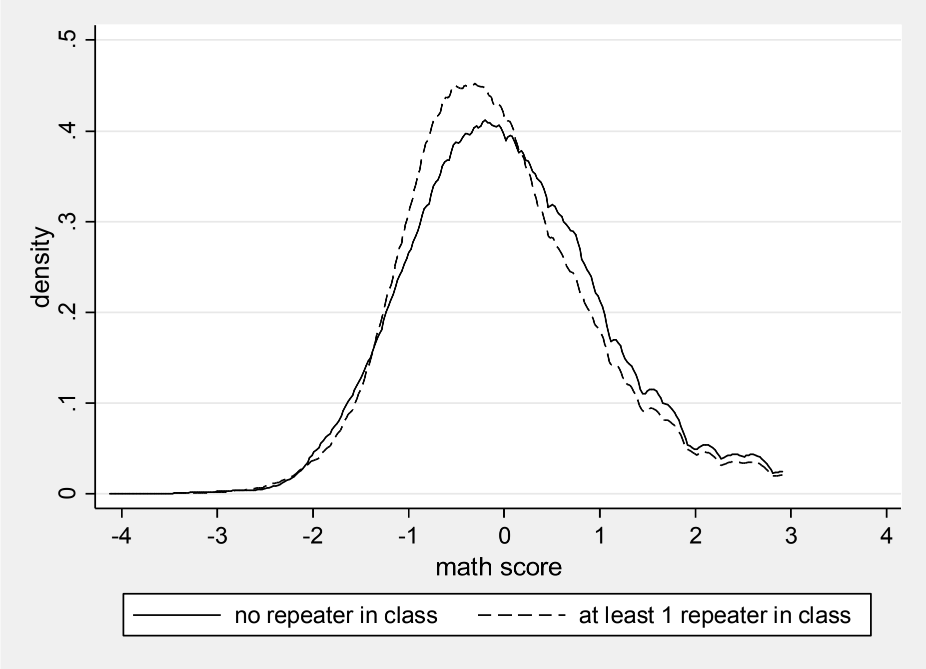

Figure 3A: Kindergarten math scores by repeater exposure

Notes: The figure plots smoothed kernel densities of end-of-kindergarten math scores separately for students with and without repeater exposure. Repeater exposure is measured as being in class with at least one repeater. The sample includes the 5,573 non-repeating students in the estimation sample for whom end-of-kindergarten math scores are available.

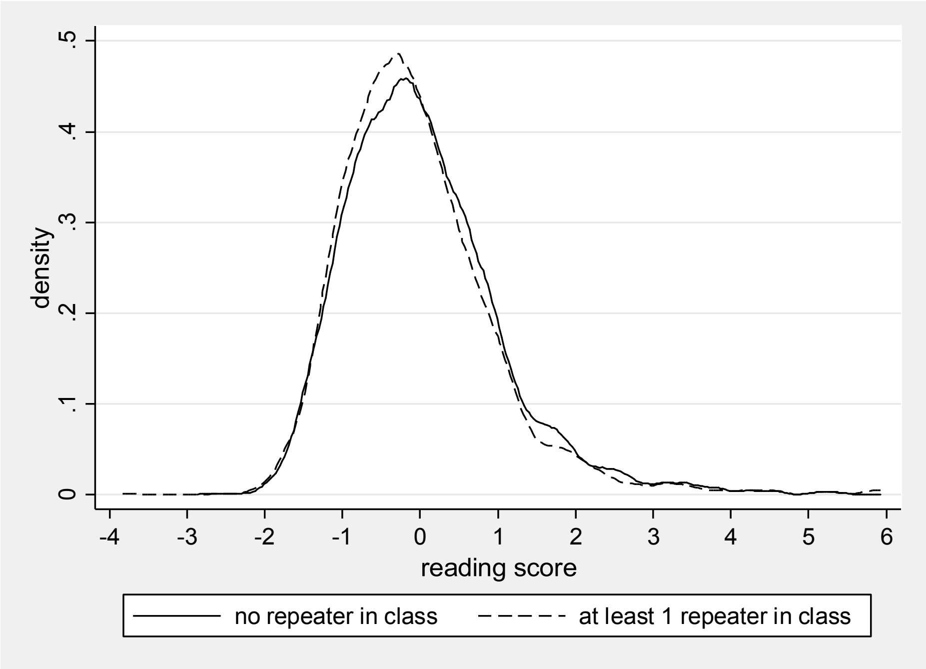

Figure 3B: Kindergarten reading scores by repeater exposure

Notes: The figure plots smoothed kernel densities of end-of-kindergarten reading scores separately for students with and without repeater exposure. Repeater exposure is measured as being in class with at least one repeater. The sample includes the 5,494 non-repeating students in the estimation sample for whom end-of-kindergarten reading scores are available.

Table 1: Descriptive statistics

	Non-repeaters			Repeaters		
	N	Mean	SD	N	Mean	SD
Demographic characteristics						
Male	6,039	0.51	0.50	193	0.70	0.46
Black	6,039	0.33	0.47	193	0.17	0.38
Free lunch	6,039	0.48	0.50	193	0.65	0.48
Age in years	6,039	5.48	0.31	193	6.39	0.31
Old for grade	6,039	0.03	0.17	193	1.00	0.00
Repeater exposure						
At least 1 repeater in class	6,039	0.39	0.49	–	–	–
Kindergarten test scores						
Math score	5,614	0.00	1.00	175	-0.36	0.80
Reading score	5,535	0.00	1.00	173	-0.47	0.69
Elementary and middle school test scores						
1st-grade math score	4,234	0.00	1.00	128	-0.72	0.91
1st-grade reading score	4,144	0.00	1.00	109	-0.80	0.73
5th-grade math score	4,479	0.00	1.00	113	-0.94	1.34
5th-grade reading score	4,481	0.00	1.00	113	-0.89	1.26
8th-grade math score	4,353	0.00	1.00	102	-0.88	1.09
8th-grade reading score	4,364	0.00	1.00	108	-0.93	1.15
Non-cognitive skills						
4th-grade effort	1,628	0.00	1.00	32	-1.13	1.24
4th-grade initiative	1,628	0.00	1.00	32	-1.01	1.01
4th-grade discipline	1,628	0.00	1.00	32	-0.32	1.20
8th-grade effort	1,731	0.00	1.00	37	-0.50	1.09
8th-grade initiative	1,731	0.00	1.00	37	-0.43	0.91
8th-grade discipline	1,731	0.00	1.00	37	-0.29	1.06
Long-term outcomes						
High school GPA	2,438	81.36	8.45	40	79.51	8.36
High school graduation	2,955	0.87	0.34	60	0.67	0.48
Took ACT/SAT	6,039	0.41	0.49	193	0.12	0.32

Notes: The table shows descriptive statistics separately for the 6,039 non-repeating students and the 193 repeaters in the estimation sample. Age in years is defined as age in days on September 30, 1985, divided by 365.25. A student is considered old for grade if based on her age and Tennessee's kindergarten entry cutoff date of September 30 she would be expected to attend at least first grade in the 1985-86 school year. Repeater exposure is not defined for repeaters because this paper studies spillovers from repeaters on non-repeating students. The non-cognitive skill measures are indices summarizing teacher ratings of student behavior in three areas: effort, initiative, and discipline. All test scores and measures of non-cognitive skills are standardized to have mean 0 and standard deviation 1 for non-repeating students. High school GPA is measured on a scale from 0-100. Took ACT/SAT is an indicator for whether the student took either of these tests in 1998, when most students were in their final year of high school.

Table 2: Kindergarten repeater exposure and end-of-kindergarten test scores

	Math (1)	Math (2)	Reading (3)	Reading (4)
Repeater exposure	-0.092 **	-0.093 **	-0.016	-0.017
	(0.044)	(0.042)	(0.046)	(0.045)
Male		-0.144 ***		-0.175 ***
		(0.024)		(0.025)
Black		-0.354 ***		-0.247 ***
		(0.051)		(0.053)
Free lunch		-0.411 ***		-0.451 ***
		(0.029)		(0.029)
Age in years		0.550 ***		0.409 ***
		(0.044)		(0.048)
Old for grade		-0.408 ***		-0.342 ***
		(0.081)		(0.075)
Class size	-0.019 ***	-0.018 ***	-0.023 ***	-0.021 ***
	(0.006)	(0.006)	(0.006)	(0.005)
Observations	5,614	5,614	5,535	5,535

Notes: The table reports estimates from regressions of end-of-kindergarten math and reading scores on the variables listed in rows and school fixed effects. Test scores are standardized to have mean 0 and standard deviation 1 across all non-repeating students in the estimation sample. Repeater exposure is measured as an indicator taking value 1 if the student's class contains at least one repeater and 0 otherwise. Regressions include all non-repeating students in the estimation sample for whom the outcome is observed. Standard errors in parentheses allow for clustering at the class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 3: Heterogeneity by demographic characteristics of non-repeating students

	Math (1)	Math (2)	Math (3)	Math (4)	Math (5)
Repeater exposure	-0.077	-0.080 *	-0.066	-0.095 **	-0.101 **
	(0.047)	(0.046)	(0.048)	(0.042)	(0.042)
Repeater exposure * male	-0.031				
	(0.047)				
Repeater exposure * black		-0.052			
		(0.079)			
Repeater exposure * free lunch			-0.066		
			(0.055)		
Repeater exposure * age in years				0.149	
				(0.092)	
Repeater exposure * old for grade				0.040	
				(0.160)	
Repeater exposure * dem. background index					0.057 *
					(0.030)
Student background	Y	Y	Y	Y	Y
Observations	5,614	5,614	5,614	5,614	5,614

Notes: The table reports estimates from regressions of the end-of-kindergarten math score on the variables listed in rows, class size, and school fixed effects. Math scores are standardized to have mean 0 and standard deviation 1 across all non-repeating students in the estimation sample. Age in years in column 4 is the demeaned version of the variable reported in Table 1. Specifications in columns 1-4 control for student background using the five demographic characteristics reported in Table 1, whereas the specification in column 5 controls for the demographic background index instead. See text for details on how the demographic background index is constructed. Repeater exposure is measured as an indicator taking value 1 if the student's class contains at least one repeater and 0 otherwise. Regressions include the 5,614 non-repeating students with an end-of-kindergarten math score in the estimation sample. Standard errors in parentheses allow for clustering at the class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 4: Kindergarten repeater exposure and elementary and middle school test scores

	1st grade		End of elementary school 5th grade		End of middle school 8th grade	
	Math	Reading	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Baseline specification						
Repeater exposure	0.023	0.006	0.043	0.034	0.059 *	0.018
	(0.037)	(0.035)	(0.032)	(0.033)	(0.034)	(0.032)
Panel B: Interaction with demographic background index						
Repeater exposure	0.006	-0.012	0.034	0.027	0.052	0.006
	(0.039)	(0.037)	(0.033)	(0.034)	(0.034)	(0.033)
Repeater exposure *	0.060 *	0.067 **	0.045	0.031	0.033	0.054 *
dem. background index	(0.033)	(0.032)	(0.033)	(0.032)	(0.032)	(0.032)
Observations	4,234	4,144	4,479	4,481	4,353	4,364

Notes: The table reports estimates from regressions of the dependent variables listed in columns on the variables listed in rows, kindergarten class size, and kindergarten school fixed effects. Test scores are standardized to have mean 0 and standard deviation 1 across all non-repeating students in the estimation sample. Specifications in panel A also control for the student demographic characteristics listed in Table 1. Specifications in panel B also control for the demographic background index. Repeater exposure is measured as an indicator taking value 1 if the student's kindergarten class contains at least one repeater and 0 otherwise. Regressions include all non-repeating students in the estimation sample for whom the outcome is observed. Standard errors in parentheses allow for clustering at the kindergarten class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 5: Kindergarten repeater exposure and non-cognitive skills

	4th grade			8th grade			
	Effort (1)	Initiative (2)	Discipline (3)	Effort (4)	Initiative (5)	Discipline (6)	Non-cognitive skills index (7)
Panel A: Baseline specification							
Repeater exposure	0.097 *	0.017	0.142 ***	0.166 ***	0.101 *	0.194 ***	0.121 ***
	(0.054)	(0.056)	(0.054)	(0.055)	(0.056)	(0.052)	(0.043)
Panel B: Interaction with demographic background index							
Repeater exposure	0.130 **	0.032	0.164 ***	0.156 ***	0.119 *	0.169 ***	0.131 ***
	(0.059)	(0.062)	(0.061)	(0.064)	(0.065)	(0.065)	(0.049)
Repeater exposure *	-0.071	-0.032	-0.047	0.011	-0.044	0.036	-0.019
dem. background index	(0.055)	(0.060)	(0.057)	(0.055)	(0.054)	(0.058)	(0.046)
Observations	1,628	1,628	1,628	1,731	1,731	1,731	2,440

Notes: The table reports estimates from regressions of the dependent variables listed in columns on the variables listed in rows, kindergarten class size, and kindergarten school fixed effects. Specifications in panel A also control for the student demographic variables listed in Table 1. Specifications in panel B also control for the demographic background index. The outcome variables in columns 1-6 are indices summarizing teacher ratings of student behavior in three areas: effort, initiative, and discipline. The indices are standardized to have mean 0 and standard deviation 1 across all non-repeating students in the estimation sample. The non-cognitive skills index used in column 7 is constructed by averaging these six standardized indices and normalizing them to have mean 0 and standard deviation 1. Repeater exposure is measured as an indicator taking value 1 if the student's kindergarten class contains at least one repeater and 0 otherwise. Regressions include all non-repeating students in the estimation sample for whom the outcome is observed. Standard errors in parentheses allow for clustering at the kindergarten class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 6: Kindergarten repeater exposure and long-term outcomes

	High school GPA (1)	High school graduation (2)	Took ACT/SAT (3)	Long-term outcomes index (4)
Panel A: Baseline specification				
Repeater exposure	0.652 *	0.023 *	0.033 **	0.078 ***
	(0.349)	(0.013)	(0.014)	(0.028)
Panel B: Interaction with demographic background index				
Repeater exposure	0.663 *	0.026 *	0.028 *	0.068 **
	(0.394)	(0.015)	(0.014)	(0.029)
Repeater exposure *	0.141	-0.008	0.036 ***	0.058 **
dem. background index	(0.400)	(0.016)	(0.014)	(0.028)
Observations	2,438	2,955	6,039	6,039

Notes: The table reports estimates from regressions of the dependent variables listed in columns on the variables listed in rows, kindergarten class size, and kindergarten school fixed effects. Specifications in panel A also control for the student demographic characteristics listed in Table 1. Specifications in panel B also control for the demographic background index. See text for details on how the long-term outcomes index is constructed. Repeater exposure is measured as an indicator taking value 1 if the student's kindergarten class contains at least one repeater and 0 otherwise. Regressions include all non-repeating students in the estimation sample for whom the outcome is observed. Standard errors in parentheses allow for clustering at the kindergarten class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 7: Kindergarten repeater exposure, non-cognitive skills, and long-term outcomes

	Long-term outcomes index					
	(1)	(2)	(3)	(4)	(5)	(6)
Repeater exposure			0.080 *	0.030	0.087 **	0.037
			(0.044)	(0.039)	(0.043)	(0.039)
Non-cognitive skills index	0.417 ***	0.382 ***		0.416 ***		0.381 ***
	(0.019)	(0.020)		(0.018)		(0.020)
Kindergarten composite score		0.152 ***			0.292 ***	0.152 ***
		(0.024)			(0.024)	(0.024)
Observations	2,440	2,440	2,440	2,440	2,440	2,440
$\chi^2(\hat{\beta}_{1,restricted}=\hat{\beta}_{1,unrestricted})$			8.05		10.22	
$Probability>\chi^2$			0.005		0.001	

Notes: The table reports estimates from regressions of the long-term outcomes index on the variables listed in rows, kindergarten class size, the five demographic characteristics reported in Table 1, and kindergarten school fixed effects. Repeater exposure is measured as an indicator taking value 1 if the student's kindergarten class contains at least one repeater and 0 otherwise. The kindergarten composite score is computed as the average of the end-of-kindergarten math and reading scores. See text for details on how the long-term outcomes index and the non-cognitive skills index are constructed. Regressions include the 2,440 non-repeating students in the estimation sample for whom both non-cognitive skills and end-of-kindergarten scores are observed. Standard errors in parentheses allow for clustering at the kindergarten class level. The chi-square statistic reported in the second from last row is for a test of the null hypothesis that the coefficients on repeater exposure in columns 3 and 4 (in columns 5 and 6) are equal. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Appendix Table 1: Randomization tests

	Male (1)	Black (2)	Free lunch (3)	Age in years (4)	Old for grade (5)
Panel A: Repeater exposure dummy					
Repeater exposure	-0.005	-0.001	0.004	0.001	-0.003
	(0.015)	(0.007)	(0.015)	(0.009)	(0.005)
Observations	6,039	6,039	6,039	6,039	6,039
Panel B: Number of repeaters in class					
Number of repeaters / 100	-0.910	-0.791 **	0.303	-0.008	-0.417
	(0.829)	(0.333)	(0.841)	(0.509)	(0.273)
Observations	6,039	6,039	6,039	6,039	6,039

Notes: The table reports estimates from regressions of the dependent variables listed in columns on the variables listed in rows and school fixed effects. Repeater exposure in panel A is measured as an indicator taking value 1 if the student's class contains at least one repeater and 0 otherwise. Specifications in panel B include the number of repeaters in a student's class as treatment instead. Regressions include all non-repeating students in the estimation sample. Standard errors in parentheses allow for clustering at the class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Appendix Table 2: Robustness to using alternative measures of repeater exposure

	Number of classes	Kindergarten math score (1)	8th-grade math score (2)	Non-cognitive skills index (3)	Long-term outcomes index (4)
Panel A: Indicators for different numbers of repeaters in class					
1 repeater in class	81	-0.097 **	0.069 *	0.125 ***	0.077 **
		(0.046)	(0.036)	(0.047)	(0.031)
2 repeaters in class	34	-0.099	0.038	0.133 **	0.093 **
		(0.070)	(0.053)	(0.068)	(0.042)
3-5 repeaters in class	13	-0.031	0.017	0.032	0.039
		(0.103)	(0.090)	(0.101)	(0.065)
Panel B: Linear share of repeaters in class					
Share of repeaters	325	-0.626	0.366	1.052 **	0.783 **
		(0.485)	(0.406)	(0.465)	(0.311)
Panel C: Indicators for different shares of repeaters in class					
Share of repeaters 0.1–5%	46	-0.109 *	0.079 *	0.160 ***	0.067 *
		(0.057)	(0.046)	(0.060)	(0.037)
Share of repeaters 5–10%	53	-0.100	0.030	0.094	0.077 **
		(0.062)	(0.043)	(0.058)	(0.038)
Share of repeaters >10%	29	-0.030	0.073	0.091	0.110 **
		(0.073)	(0.065)	(0.071)	(0.044)

Notes: The table reports estimates from regressions of the dependent variables listed in columns on the variables listed in rows, controls for class size and the student demographic characteristics listed in Table 1, and kindergarten school fixed effects. Repeater exposure is measured by indicators for whether a class contains 1, 2, or 3-5 repeaters in panel A, by the share of repeating students in class in panel B, and by indicators for whether 0.1-5%, 5-10%, or more than 10% of students in a class are repeaters in panel C. The second from left column reports the number of classes in the estimation sample that contain the indicated number or share of repeaters. See text for details on how the outcome variables in these regressions are constructed. Regressions are run for all non-repeating students in the estimation sample for whom the outcome is observed. Standard errors in parentheses allow for clustering at the kindergarten class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Appendix Table 3: Robustness to various sample restrictions

	Kindergarten math score (1)	8th-grade math score (2)	Non-cognitive skills index (3)	Long-term outcomes index (4)
Panel A: Only regular-sized classes				
Repeater exposure	-0.081 *	0.089 **	0.203 ***	0.041
	(0.045)	(0.042)	(0.056)	(0.034)
Observations	3,921	3,034	1,633	4,221
Panel B: Only schools with positive repeater shares				
Repeater exposure	-0.094 **	0.053	0.126 ***	0.075 ***
	(0.042)	(0.034)	(0.043)	(0.028)
Observations	4,334	3,421	1,955	4,654
Panel C: Excluding the school with more than 18% of repeaters				
Repeater exposure	-0.093 **	0.060 *	0.122 ***	0.078 ***
	(0.042)	(0.034)	(0.043)	(0.028)
Observations	5,573	4,317	2,428	5,996

Notes: The table reports estimates from regressions of the dependent variables listed in columns on the variables listed in rows, the student demographic characteristics listed in Table 1, and kindergarten school fixed effects. Specifications in panels B and C additionally control for class size. In panel A, the sample is restricted to non-repeating students in regular-sized kindergarten classes. In panel B, the sample is restricted to non-repeating students in schools that contain at least one kindergarten repeater. In panel C, 53 students in the school in which more than 18% of kindergarten students are repeaters are dropped from the sample. See text for details on how the outcome variables in these regressions are constructed. Standard errors in parentheses allow for clustering at the kindergarten class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Appendix Table 4: Tests for selective attrition from the sample by kindergarten repeater exposure

	Kindergarten math score (1)	1st-grade math score (2)	5th-grade math score (3)	8th-grade math score (4)	4th-grade non-cognitive skills (5)	8th-grade non-cognitive skills (6)	High school GPA (7)	High school graduation (8)
	Panel A: Outcome is an indicator for being observed with the variable in the column head							
Repeater exposure	-0.010	0.030 **	-0.004	0.009	-0.011	-0.019	0.011	0.008
	(0.008)	(0.013)	(0.013)	(0.012)	(0.015)	(0.013)	(0.013)	(0.014)
	Panel B: Outcome is an indicator for being observed with the variable in the column head							
Repeater exposure	-0.010	0.030 **	-0.006	0.007	-0.015	-0.020	0.008	0.005
	(0.008)	(0.013)	(0.013)	(0.012)	(0.015)	(0.013)	(0.013)	(0.014)
Repeater exposure *	-0.003	0.002	0.019	0.019	0.024 *	0.005	0.022	0.017
dem. background index	(0.007)	(0.013)	(0.013)	(0.013)	(0.013)	(0.015)	(0.014)	(0.014)
	Panel C: Outcome is the end-of-kindergarten math score, sample is restricted to students observed with the variable in the column head							
Repeater exposure	-0.093 **	-0.109 **	-0.100 **	-0.090 **	-0.075	-0.034	-0.063	-0.130 **
	(0.041)	(0.044)	(0.044)	(0.045)	(0.061)	(0.057)	(0.055)	(0.051)
Observations	5,614	3,987	4,216	4,105	1,544	1,627	2,317	2,794
	Panel D: Outcome is the ACT/SAT test taking dummy, sample is restricted to students observed with the variable in the column head							
Repeater exposure	0.038 **	0.042 **	0.033 **	0.029 *	0.016	0.049 **	0.043 **	0.025
	(0.015)	(0.018)	(0.015)	(0.016)	(0.030)	(0.024)	(0.020)	(0.018)
Observations	5,614	4,234	4,479	4,353	1,628	1,731	2,438	2,955

Notes: The table reports estimates from a series of tests for selective attrition by kindergarten repeater exposure. In panels A and B, the dependent variable is an indicator taking value 1 if the outcome in the column head is observed for a given student and 0 otherwise. Regressions in these two panels are based on the sample of 6,039 non-repeating students. Specifications in panel A also control for kindergarten class size, the five demographic variables reported in Table 1, and kindergarten school fixed effects. Specifications in panel B also control for a kindergarten class size, the demographic background index, and kindergarten school fixed effects. Specifications in panels C re-estimate the regression in column 2 of Table 2 for students with an end-of-kindergarten math score and for whom the outcome in the column head is observed. Specifications in panel D re-estimate the regressions in column 3 of Table 6 (panel A) for students for whom the outcome in the column head is observed. Repeater exposure is measured by a dummy taking value 1 if the student's kindergarten class contains at least one repeater and 0 otherwise. Standard errors in parentheses allow for clustering at the kindergarten class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Appendix Table 5: Robustness to relative (to repeaters) measurement of non-cognitive skills

	4^{th}-grade effort (1)	4^{th}-grade initiative (2)	4^{th}-grade discipline (3)
Repeater exposure	0.069	-0.014	0.112 **
	(0.056)	(0.060)	(0.055)
Observations	1,445	1,445	1,445

Notes: The table reports estimates from regressions of the dependent variables listed in columns on the repeater-exposure dummy, kindergarten class size, the student demographic variables listed in Table 1, and kindergarten school fixed effects. The sample in these regressions is restricted to students who are in fourth-grade classes that do not contain any of the 193 initially observed kindergarten repeaters. Repeater exposure is measured as an indicator taking value 1 if the student's kindergarten class contains at least one repeater and 0 otherwise. See text for definition of the dependent variables. Standard errors in parentheses allow for clustering at the kindergarten class level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Appendix Table 6: Spillovers from male versus female repeaters

	Kindergarten math score (1)	Kindergarten reading score (2)
Male repeater in class	-0.167 ***	-0.070
	(0.047)	(0.045)
Female repeater in class	-0.031	0.024
	(0.078)	(0.085)
H0: Male = female (*p*-value)	0.12	0.27
Observations	5,210	5,134

	Difference between male and female repeaters
Kindergarten math score	0.372*
	(0.188)
Kindergarten reading score	0.195
	(0.190)
Fourth grade discipline	-2.013*
	(1.092)

Notes: The upper part of the table reports estimates from regressions of end-of-kindergarten math and reading scores on indicators for being in class with a male or a female repeater, class size, the student demographic characteristics listed in Table 1, and kindergarten school fixed effects. The sample in these regressions is restricted to non-repeating students in classes that do not contain both male and female repeaters. The reported *p*-value in each column is for a test of equivalent impacts from male and female repeater exposure. The lower part of the table reports coefficients from regressions of the dependent variables listed in rows on a male indicator and kindergarten school fixed effects. The sample is restricted to repeaters in these regressions. Standard errors in parentheses allow for clustering at the kindergarten class level in the upper part of the table and at the kindergarten school level in the lower part of the table. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.