Parental Education Spillovers and School Outcomes: Evidence from Siblings

Latest Version: Here

Gonçalo Lima Miguel Estevinho Nunes *

February 15, 2025

Abstract

We study the effect of exposure to peers with college-educated parents on the school outcomes of children from less educated households. For identification, we exploit within-family-by-year variation to control for the significant non-random sorting of parents into schools, and time-varying shocks to the family environment. We use a rich administrative data from Portugal, a school system with some of the largest gaps in educational attainment by children's socioeconomic status among OECD countries. We find that moving a student from the 10th to the 90th percentile in the distribution of exposure to children with college-educated parents is equivalent to closing about a fifth of the gap in grade repetition by parental education. We show that the effect on grade repetition is partly driven by improved school performance, consistent with a social contagion type of mechanism. However, we also show suggestive evidence that schools fail fewer students who were eligible to repeat when faced with higher concentrations of students with college-educated parents.

JEL Classification: I21, I24. **Keywords**: Parental education; spillovers; peer effects; grade repetition; student achievement.

Word Count: 9,444

*Gonçalo Lima: goncalo.lima@eui.eu, PhD Candidate at the Department of Economics of the European University Institute, Research Fellow at the University of Bologna. Miguel Nunes: miguel.nunes@eui.eu, PhD Candidate at the Department of Economics of the European University Institute. We are thankful to Andrea Ichino, Fabrizia Mealli, Ellen Greaves, Elias Einiö, Gustave Kenedi, Marta Curull-Sentís, Richard Murphy, as well as participants at EALE 2024, IWAEE 2024, EAYE 2024, SOLE 2024, the 3rd Meeting of Junior Economists in Milan, and seminars in the European University Institute for helpful discussions. Without the administrative data provided by the Portuguese Directorate General for Education and Science Statistics (DGEEC), under the purview of the Ministry of Education, this work would not have been possible. We are also grateful to the Economics of Education Knowledge Center, at Nova School of Business and Economics, in Portugal, for access to the data. This work was supported by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019, PTDC/EGE-ECO/4764/2021, Social Sciences DataLab, Project 22209, as well as individual PhD grants SFRH/BD/150668/2020 and PRT/BD/153679/2022), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209). All views and errors are ours.

1. Introduction

College-educated parents set greater educational expectations, provide more stimulating home environments, and invest more in their children's development compared to less educated parents (Müller, 2023; Boneva and Rauh, 2018; Guryan et al., 2008; Bianchi and Robinson, 1997; Hill and Stafford, 1974). However, it remains unclear to what extent children of less educated parents benefit from exposure to more educated families. Schools provide a unique environment where social norms, behaviors, and educational expectations characteristic of college-educated households can influence other students through social contagion. On the other hand, schools may respond to more children with college-educated parents in ways that could ambiguously impact the learning outcomes of other students.

Identifying the effect of exposure to highly educated parents is complicated by two main empirical challenges. First, children of highly educated parents are not randomly assigned to schools. Income-based residential segregation—highly correlated to parental education—largely explains sorting across educational institutions (Owens et al., 2016; Nechyba, 2006). Second, less educated parents may strategically enroll their children in schools with higher concentrations of college-educated parents or other advantaged peers. If these children would have achieved similar outcomes in less exposed schools, failing to account for this non-random sorting could lead to an overestimation of potential benefits. Conversely, more highly educated parents may decide to leave schools as a response to increased integration of children from disadvantaged backgrounds (Cascio and Lewis, 2012; Hoxby, 2000).

In settings without random assignment, existing research often relies on variation in student composition across cohorts within the same school to identify the effect of peer characteristics on school outcomes (Barrios-Fernández, 2023; Epple and Romano, 2011). Interpreting these effects as causal hinges on the assumption that parents sort across schools based on the past composition of student cohorts, rather than the current or anticipated future cohort compositions. While cross-cohort variation in characteristics like gender is arguably easier to be taken as idiosyncratic (e.g. Lavy and Schlosser, 2011; Hoxby, 2000), the same assumption is more difficult to justify for characteristics strongly correlated to school sorting patterns, such as parental background.

In this paper, we tackle parental sorting by exploiting idiosyncratic variation within the family. We examine how exposure to peers with college-educated parents affects the educational attainment of children without college-educated parents, leveraging variation across siblings who are exposed to different cohort compositions. We study this question in the context of Portugal, for which we have detailed administrative data on the universe of students enrolled in public primary and lower secondary schools. We focus on grade repetition as our measure of educational attainment due to its consistent availability in the data, as well as for its relevance for students' educational and career paths in this setting.

Studying this question in the context of the Portuguese education system is insightful for mainly three reasons. First, grade repetition is quite prevalent, with Portugal having one of the highest retention

rates among OECD countries (OECD, 2022). In our data, more than a quarter of students repeat a grade at least once before high school. Second, Portugal has the second-largest grade repetition gap by parental background among high-income countries (OECD, 2020). In our data, children without college-educated parents are six times more likely to repeat a grade than those with at least one parent who completed college. Empirical evidence shows that, while grade repetition may remedy insufficient learning through some short-term gains (Figlio and Özek, 2020; Schwerdt et al., 2017), it has substantial negative effects on earnings in the labor market (ter Meulen, 2023). Therefore, grade repetition is not without relevant economic tradeoffs for individuals and public policy. Third, there is significant segregation across schools based on parental education, even within the public school system. Children with college-educated parents—just a quarter of the population in our period of analysis—have about twice as many classmates whose parents completed college compared to those without college-educated parents. Strong patterns of residential sorting contribute to Portugal having a level of socioeconomic segregation across schools similar to that of the largest metropolitan areas in the US (Liebowitz et al., 2018; Owens et al., 2016).

Portugal's rich data helps us address our main empirical challenges. First, it enables us to measure the size and composition of student cohorts across schools, grades and years. We define as an individual's peers in a given year all other students enrolled in the same grade cohort, rather than just those belonging to the same classroom. As such, we recognize the potential biases introduced by empirically verified non-random allocation of students to classrooms within schools. In contrast to selective allocation of students to classes, it is unlikely that schools are able to exert any relevant discretion in sorting students across grade cohorts.

Second, we are able to construct multiple measures of peer exposure. Our main variable of interest is cumulative exposure to peers with college-educated (CE) parents. Relying on the panel structure of the data, we compute the average exposure for each individual across years—from first grade to any grade in which they are observed. Thus, for each child-year cell we have a continuous measure of intensity of exposure to CE peers throughout the individual's schooling path until then. To test different hypotheses and ensure robustness, we also present analyses with alternative definitions of exposure.

Third, we rely on an anonymized family identifier to implement our research design. For estimation, we use family-by-year fixed effects, which allow us to control for all family background characteristics in each year. To control for differences across time, schools and grades we also include school-by-year and grade-by-year fixed effects in our preferred specifications. Thus, the identifying variation comes from siblings enrolled in different grades in the same school and academic year, and siblings enrolled in the same grade but in different schools during the same academic year. Our main identifying assumption is that these idiosyncratic differences in exposure within the family are as good as random. We provide evidence in support of our empirical strategy, and run a series of alternative specifications to attest its robustness.

We find that exposure to students with college-educated parents significantly reduces the likelihood of grade repetition among children from less educated households. For an average cohort of 80 students, increasing exposure from 2 peers with CE parents (10th percentile) to 30 peers (90th percentile) decreases grade repetition by about 1 percentage point, or 13% of the sample mean. The magnitude of this change corresponds to about one-fifth of the gap in grade repetition between children with and without CE parents. We also find that average exposure in previous grades affects academic achievement beyond the effect of contemporaneous exposure. The qualitative interpretation of our results remains robust across different definitions of exposure and specifications.

Importantly, our identification strategy uncovers an effect size that is just one-quarter of that found using within-school, across-cohort variation. There are two main reasons for this difference between research designs. First, each strategy uses a different source of identifying variation. Within-school, across-cohort designs compare individuals who differ along many dimensions, including in their exposure to given peers. With our research design, we restrict these differences across cohort to within-family comparisons. We argue that siblings serve as a more plausible counterfactual of each individual than just any other children in adjacent cohorts. Moreover, the concern about parental sorting is directly addressed by design. Our estimator is thus relatively more conservative in the type of comparisons it allows. Second, each strategy contends with different types of contaminating spillover effects. With our research design, we can explicitly infer that the magnitude of the estimates may be attenuated by within-family spillovers. Thus, we carefully address how sibling spillovers may bias our results. Relying on minimal assumptions and supporting evidence on the monotonicity of these intrahousehold spillovers, we argue that the estimated effects are lower bounds of our estimand of interest. On the other hand, with within-school, across-cohort designs potential sources of interference are often more difficult to identify and seldom discussed. Throughout the paper, we interpret each of our results in light of previous evidence and the corresponding source of identifying variation.

To test the interpretation of our findings as truly stemming from parental education, we characterize families with college-educated parents. For instance, children with CE parents are significantly less likely to be born to poor and immigrant families. Reassuringly, we find that peers' parental education still matters for educational attainment beyond these other family characteristics.

We also investigate how the effects vary across different types of families. We test for effect heterogeneity in terms of family size, poverty status, and siblings gender mix. Consistent with an hypothesis of stronger influence of school inputs when parental investment is distributed over more children, students from larger families benefit differentially more from being exposed to peers with CE parents. On the other hand, relatively poorer children benefit less from higher exposure to peers with CE parents. Therefore, children from poor families are doubly penalized, as they both benefit less from higher exposure and are significantly less likely to being exposed to peers with CE parents. Lastly, we find that the

estimated effects are similar across only-male and mixed-gender sibling groups, while finding no effect for families with only daughters.

Our findings are reduced-form estimates of the effect of exposure on school outcomes. Although we can observe exposure to children with CE parents throughout time, we cannot identify whether children from less educated households actually engage in meaningful relationships with the former, and whether this is the main channel through which exposure affects school outcomes. The effects of exposure to children of more educated parents may be beneficial through other channels other than the strict contact between students of these different subgroups, also depending on how schools react to different levels of exposure.

Relevantly, we find evidence that the reduction in grade repetition is likely driven by improved school performance. With the caveat of restricting this analysis to middle school siblings—for which we have performance information—we investigate how exposure to peers with CE parents affects academic outcomes upstream of grade repetition. We find that moving a student from the 10th to the 90th percentile of cumulative exposure increases GPA by 6.5% of a standard deviation. Additionally, more exposed siblings fail at significantly fewer subjects, including both language and math. These findings support the hypothesis that the effect on repetition hinges on effective learning gains, rather than just changes in school policy regarding student retention.

However, we find suggestive evidence that the discretion with which schools repeat students can also explain part of the estimated effects. Only about 60% of the students that are eligible to repeat the grade according to set guidelines are ultimately failed. Schools are more or less lenient in terms of their grade repetition policies, both across time and across cohorts. We show that higher concentrations of children with college-educated parents within the same school, leads to fewer children without college-educated parents repeating, even if eligible.

Related literature.— Our paper makes empirical contributions to mainly three strands of literature. First, it adds to the large body of evidence on peer effects in educational outcomes (surveyed in Barrios-Fernández, 2023; Epple and Romano, 2011; Sacerdote, 2011). Extant research shows that students tend to benefit from exposure to higher ability peers in the classroom, or to peer characteristics that predict higher school achievement. To the best of our knowledge, ours is the first paper to look at the effect on the likelihood of being retained in the same grade, a relevant outcome for school systems that make extensive use of grade repetition as a potential remedy for poor student performance.

Second, it contributes to the growing literature on the effects of peers' parental education on school outcomes. Empirical evidence on this particular type of peer effects is mixed. In the US, being assigned to a kindergarten class with greater exposure to highly-educated parents leads to higher achievement in math and reading, but not on socioemotional skills (Fruehwirth and Gagete-Miranda, 2019). Likewise, mothers' mean education in class is an important determinant of eight grade achievement in Chile

(McEwan, 2003). Relatedly, Bonesronning and Haraldsvik (2014) find a negative spillover effect of peers from less-educated households on fifth grade achievement in Norway. On the other hand, Bifulco et al. (2011) do not find significant spillover effects of parental education on final high school GPA in the US. In Denmark, Bertoni et al. (2020) find sizeable positive effects on the likelihood of college graduation and higher earnings in the labor market. Exploiting heterogeneity on peers' gender, Cools et al. (2019) find that girls exposed to higher concentrations of boys with highly educated parents have lower self-esteem and aspirations, ultimately reducing their probability of completing a bachelor's degree. Close to our research question, Cattan et al. (2023) find that students whose parents did not attend elite colleges are more likely to attend these colleges when more exposed to peers with such family background. Contrary to previous evidence, our paper goes beyond differences in contemporaneous exposure, exploiting the average cumulative exposure since children start school, identifying the effects across multiple grades, in both primary and lower education.

Finally, this paper contributes to existing empirical evidence exploiting within-family comparisons as identifying source of variation (e.g. Figlio et al., 2023a; Bertoni et al., 2020; Bonesronning and Haraldsvik, 2014). Observational studies on parental education spillovers have mostly exploited variation in student composition across cohorts within schools (as in Hoxby, 2000) to identify the effects on outcomes (Cattan et al., 2023; Cools et al., 2019; Fruehwirth and Gagete-Miranda, 2019; Bifulco et al., 2011; McEwan, 2003). In the context of immigrant peer effects, Figlio et al. (2023a) show that failing to partial out unobserved non-random selection of families into schools would change the qualitative interpretation of the effects of exposure to immigrants on the achievement of US-born students. Our paper contributes to this literature by following a similar strategy to the one in Figlio et al. (2023a), exploiting differences across siblings exposed to different shares of children with highly educated parents. Our paper is also close to Bertoni et al. (2020), which exploits variation in exposure across siblings attending the same school at different points in time. To the best of our knowledge, ours is the first paper to exploit within-family-by-year variation for identifying spillover effects of parental education. Moreover, we depart from these previous papers by carefully addressing potential spillovers across siblings. Given the growing empirical evidence on the effect of older siblings on the educational choices and outcomes of their siblings (e.g. Figlio et al., 2023b; Altmejd et al., 2021), we provide evidence that within-family designs provide a lower bound of the parental education spillovers running through the children's cohort peers.

Outline— The paper is structured as follows. Section 2 presents the data sources, main variables of interest, and motivating descriptive evidence. Section 3 describes our identification strategy and discusses the empirical validity of our approach. Section 4 presents the main effects on grade repetition throughout primary and lower secondary education, heterogeneity analysis as well as robustness checks to the main estimates. Section 5 documents and discusses suggestive evidence on some mechanisms explaining our results. Section 6 concludes.

2. Data and Setting

In this section we present the data we use in the empirical analysis, as well as descriptive and motivating evidence for our setting. Our analysis covers a sub-sample of 200,000 individuals who have at least a sibling enrolled in a school in the same academic year, as required by our research design (described in Section 3).[1] We use the complete dataset to present the main stylized facts in Section 2.4. We describe our main analysis sample in Section 2.5.

2.1. Data Sources

We use a de-identified dataset combining administrative school records maintained by Portugal's Ministry of Education.[2] This dataset includes all students enrolled in primary, lower, and upper secondary education in mainland Portugal (Grades 1 through 12) from 2006 to 2018. Each observation contains a unique student identifier, enabling us to track students across schools and academic years. The dataset also contains detailed demographic information, academic achievement, and attainment of each student. Crucially, we have data on parents' (or legal guardians') education levels. Because this information is missing for most private school students, we restrict our analysis to public school students, who make up 85% of the total across all grades. Additionally, we make use of an anonymous family identifier, allowing us to exploit within-family variation.

2.2. Main Variables

Exposure to Children with College-Educated Parents.— We start by identifying students with at least one college-educated parent (holding a bachelor's degree or higher). We define as having college-educated parents all individuals for which at least one of the parents has completed a bachelor degree. If information is missing for one parent, we use information on the education level of the other parent. We then compute a cumulative measure of exposure to peers with college-educated parents, beginning in the first year a student is observed in first grade ($\underline{t}$). For each enrolled student, we compute the share of peers in the same cohort—defined by school, grade, and academic year—whose parents are college-educated.[3] Thus, for each student i, in school s, grade g, and year t:

$$E_{isgt} = \frac{1}{t - \underline{t} + 1} \sum_{t'=\underline{t}}^{t} \frac{\#\text{ Peers with CE Parents}_{isgt'}}{\#\text{ Peers}_{isgt'}}. \tag{1}$$

We count the number of peers in the same cohort, rather than the same classroom. This cohort-level measure allays concerns about the endogeneity of each school's allocation of students across classes.[4]

[1] The complete dataset includes about 950,000 students enrolled in over 7,000 public schools, spanning nine different grades, between 2006 and 2018.

[2] The administrative datasets include detailed reports from public schools (MISI), private independent schools (INQ-PRIV), and publicly-subsidized private schools (MISI-PRIV). Appendix A details the data access procedure and information contained within each database.

[3] This measure of cumulative exposure is similar to the one used for exposure to foreign-born students in Figlio et al. (2023a).

[4] We empirically reject that the allocation of children across classrooms in the same cohort is random. See Appendix D for

We also explore alternative versions of exposure. First, we compute a version excluding students with unknown parental education from the denominator in Equation 1.[5] Second, we derive a measure of contemporaneous exposure, excluding prior years' exposure. Third, we create a cumulative exposure measure with a varying decay parameter, adjusting the weight of previous' years exposure (as in Figlio et al., 2023a).[6] Fourth, we compute the same measure as in Equation 1 using only peers in the same classroom. Finally, we create a version of exposure with just the average number—rather than proportion—of peers with CE parents. Additionally, in some specifications we include analogous measures of cumulative exposure to immigrants and students eligible free or reduced-price lunches (FRPL), a proxy for economic distress.

Main Outcome: Grade Repetition.— We define grade repetition by recording whether a student is enrolled in the same grade the following academic year. Typically, students are eligible to repeat a grade if they fail both Math and Language in end-of-cycle grades (4, 6 and 9) or fail three or more school subjects overall, as assessed by their teachers. However, not all eligible students repeat, as schools have some discretion in deciding whether retention would be the appropriate remedial learning strategy. Only about 59% of eligible students in lower secondary education actually repeat the grade. In Section 2.4 we provide further details and motivate why this is a relevant outcome in our setting.

Other Outcomes.— In other analysis we have additional individual-level outcomes such as students' grade point averages (GPA) in core subjects and the number and proportion of subjects failed. We also create indicators for failing Language, Math, or both subjects, given their relevance for grade repetition. Unlike the main outcome, these intermediate outcomes are only available for Grades 5 through 9.

Covariates.— Most specifications include the following covariates: Indicator variables identifying whether the student is female, foreign-born, has internet at home, whether is eligible for FRPL, and birth order. We further characterize children by age and birth spacing among siblings.

2.3. Sample Restrictions

Given our research design (described in Section 4.1), we impose multiple sample restrictions to the data. First, we start with a dataset with records of non-adult, regular education students enrolled in public schools, between first and ninth grade, for the period 2006-2018. At this stage, we compute all exposure measures described in Section 2.2. Second, we restrict the sample to individuals that we can follow since they have started school, in first grade. At this stage, we compute all outcome measures. In Appendix Table A.1 we report summary statistics for these students. Third, we further restrict observations to students that have at least a sibling who is also observed in school in our period of analysis. In Appendix Table A.2 we report summary statistics for this sample. Children with siblings are not much differentially

further details.

[5] Students with missing parental education are just 3% of the sample.

[6] See Section 4.2 for details.

selected (see Appendix Table A.3), being only slightly more likely to have an immigration background. Fourth, because we rely on family-by-year fixed effects, we restrict the sample to those siblings that can be simultaneously observed in school in the same academic year, even if in different grades (see Appendix Table A.4). Relative to the sample with all siblings, restricting to those that are observed in the same year does not produce appreciable differences in observable characteristics, except for the share of college-educated parents, which is relatively higher in the restricted sample (Appendix Table A.5). Finally, since we focus on students with no college-educated parents, we drop all students with college-educated parents (26%) or missing information on this variable (3%). Our analysis sample is composed of children without college-educated parents, who have at least a sibling enrolled in a public school in the same year, an unbalanced panel of 200,457 individuals and 766,054 observations.

2.4. Setting - Stylized Facts

In this section we provide descriptive evidence and outline the institutional background of our study, documenting four major stylized facts that help motivate our empirical analysis in the context of Portugal.

Fact 1. *Grade repetition is highly prevalent.*

Grade repetition remains a common strategy for addressing poor student performance in our setting. Portugal has the fifth highest repetition rate among 35 OECD countries: 27% of 15-year olds have repeated a grade at least once throughout their school career, delaying their entry into post-secondary education and the labor market (OECD, 2022). Although the learning effects of grade repetition are outside the scope of our analysis, it is worth noting that delaying students' grade progression is not without important costs for individuals and education systems. While extant research shows that grade repetition may benefit student learning in the short run (Jacob and Lefgren, 2004; Schwerdt et al., 2017; Figlio and Özek, 2020; Borghesan et al., 2022) it also leads to a lower likelihood of high school completion (Jacob and Lefgren, 2009; Manacorda, 2012) and reduced earnings in the labor market (ter Meulen, 2023).

In Portugal, as in other countries, grade repetition is tied to student performance in school subjects. Eligibility for grade repetition is determined by one of two necessary conditions.[7] To be eligible, students must either (i) fail at three or more school subjects; or (ii) fail at both Math and Portuguese Language in end-of-cycle grades—fourth, sixth and ninth grade. However, these conditions are not sufficient to repeat a grade, as not all eligible students repeat. Appendix Table E.1 shows that eligibility increases the likelihood of grade repetition by 59 percentage points (p.p.), a finding robust to controlling for students' individual characteristics.[8] In contrast, only 0.1% of ineligible students repeat the grade. This suggests that schools comply with necessary conditions, but exert considerable discretion on which students to

[7] Necessary conditions for grade repetition, and guidelines on the application of this type of policy in schools, are centrally set by the Ministry of Education. See, for instance, Ordinance Nr. 223-A/2018, Article 32, 6.

[8] We restrict the analysis to this subsample, because the data lacks detailed information on performance among primary school students.

repeat, conditional on eligibility.[9].

Fact 2. *Wide gaps in grade repetition and student performance by parental education.*

In most countries, socioeconomically disadvantaged students have significant lower school performance and repeat at higher rates. In Portugal, socioeconomic status (SES) is a particularly strong predictor of grade repetition, being the second OECD country in which SES most predicts grade repetition, even after controlling for differences in ability (OECD, 2020). During our period of analysis, grade repetition rates are 6% per year among students without college-educated parents, compared to just 1% among children of more educated households (Appendix Table E.2). Children without any college-educated parent are thus six times more likely to repeat the grade relative to those with at least one parent with a college degree.

Grade repetition gaps by parental education vary across years and grades. Figure 1, Panel A, shows the trend in this gap for three end-of-cycle grades—4, 6, and 9. The gap, represented by the coefficient from a regression of grade repetition on a dummy variable for having at least one college-educated parent, decreases over time. However, students from more educated households consistently repeat at lower rates. Students from more educated households repeat at relatively lower rates, although the grade repetition gap falls throughout the years. In our empirical analysis, we interact grade with academic year dummies to partial out differences across years and grades.

Like grade repetition, school performance—as measured by grade point average (GPA)—differs significantly across subgroups. Students without college-educated parents have substantially lower GPAs in lower secondary education (Appendix Table E.2), with a gap equivalent to 85% of a standard deviation. Interestingly, while these differences in school performance help explain the gap on grade repetition, we find that schools seem less favorable to students from lower educated households, conditional on being eligible for repetition. Children with CE parents are about 4 p.p. less likely to repeat than equally eligible students without CE parents (Appendix Table E.1, Columns 3 and 4).

Unsurprisingly, parental education also strongly predicts high school track choices at the end of lower secondary education. Children of college-educated parents are 22p.p. more likely to enroll in the high school track that typically leads to college education. Even after controlling for the number of grade repetitions throughout students' school paths, the gap in high school track enrollment by parental education remains large (Appendix Figure E.1).

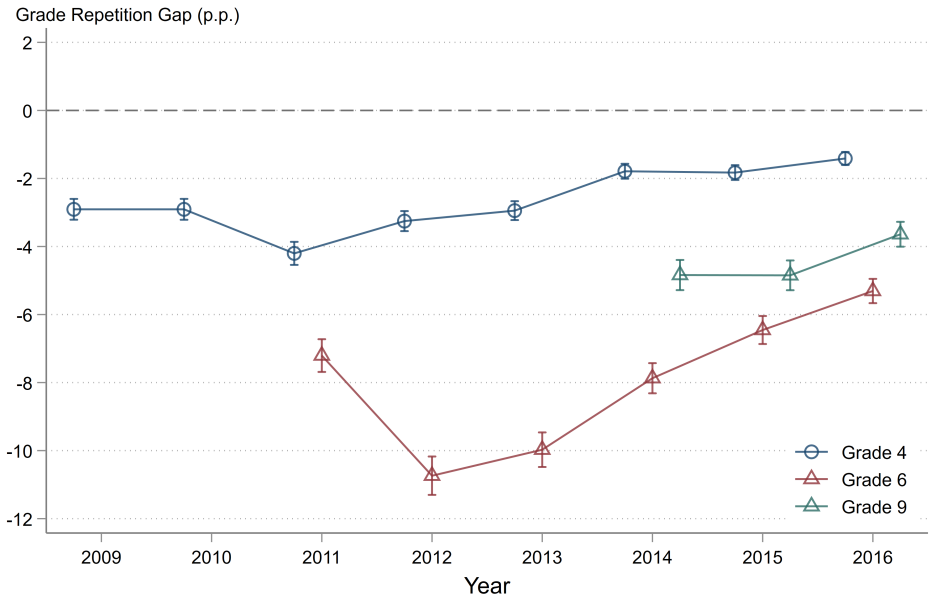
Fact 3. *Substantial segregation across schools by parental education.*

Reflecting wide disparities in parental education by neighborhood, students are highly segregated across schools. Figure 1, Panel B, depicts the distributions of cumulative exposure to CE parents by parental background. On average, children without CE parents exposed to 12.4p.p. fewer peers who

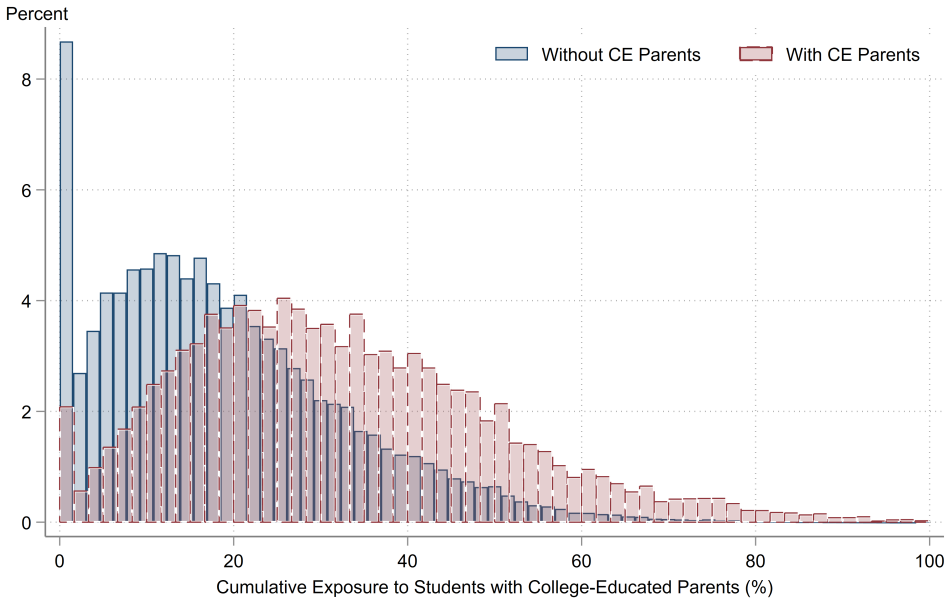
[9] In this empirical analysis, we include school-by-year fixed effects to account for differences in grade repetition policies across schools and within schools over time.

Figure 1. Grade Repetition and Cumulative Exposure to CE Parents by Parental Background

(a) Parental Education Grade Repetition Gap by Year



(b) Distributions of Cumulative Exposure to CE Parents



Notes. Panel A presents grade repetition gaps by academic year in the sample, for end-of-cycle grades (4,6, and 9), where the grade repetition gap is measured as the coefficient of a regression of grade repetition on a dummy indicating whether at least one of the parents is college-educated. Panel B depicts the distributions of the cumulative exposure variable defined in Equation 1, separately by children with and without college-educated parents. The sample includes all public school students in mainland Portugal that can be followed since Grade 1, between 2006 and 2016.

have CE parents than those with CE parents (Appendix Table E.2). Beyond the wide gap across groups, there is considerable variation in exposure within groups. About 9% of children without CE parents were never exposed to peers with CE parents, in any given year.[10] On the other hand, about a quarter of the

[10] Most observations with little exposure occur in relatively small primary schools, with an average of 15 students in the

students in each given year have cumulative exposures above 26%. [11]

Exposure to peers with college-educated parents reflects wide disparities in the distribution of these households across the territory (Appendix Figure A.4). Likewise, within municipalities, there is considerable segregation by parental education across schools (Appendix Figure A.5).

Fact 4. *Children who are more exposed to peers with college-educated parents are less likely to repeat.*

The relationship between grade repetition and cumulative exposure to peers with CE parents is downward-sloping and linear across most of its support (Appendix Figure A.3), indicating that students with greater exposure are also less likely to repeat. Importantly, this correlation masks substantial sorting across neighborhoods and schools, even if only among the individual without CE parents. Students living in poorer neighborhoods are simultaneously exposed to fewer colleagues with CE parents and at a higher risk of grade repetition (Appendix Figure A.4). In Section 4.1 we present the empirical strategy we use to uncover the effect of this type of exposure on school outcomes.

2.5. Summary Statistics

Table 1 presents summary statistics of our main analysis sample. The average child is ten years old, and less than one in every five of their colleagues in the same grade have at least a parent that has completed college. Students in our sample are relatively poor, with almost six in every ten benefiting from free or reduced price lunch benefits.[12] A small percentage (2%) of students are foreign-born, although the cumulative exposure to immigrants is almost 4%, reflecting high immigrant concentration in some neighborhoods. The sample is relatively balanced in terms of gender, with boys being slightly over-represented. The rate at which students repeat, in each year, is 7.7%. Importantly, though, 26.5% of students in the analysis sample repeat at least once while observed (Appendix Table A.7).

3. Empirical Strategy and Estimation

In this section, we present our main identifying assumptions and empirical strategy. We estimate the effect of interest by comparing siblings in the same academic year using a within-family-by-year fixed effects estimator. We also discuss evidence supporting our design, characterize the identifying variation, and address potential endogeneity concerns.

3.1. Identifying Variation

Exposure to peer characteristics is endogenous to parental decisions. Families sort into neighborhoods and enroll their children in local schools based on factors like neighborhood quality and homophily. Parents may also respond to changes in cohort composition.

grade. Appendix Figure A.1 depicts the distribution of the percentage of students with CE parents at the school-year level, separately by grade.

[11] Appendix Figure A.4 depicts the geographical distribution of student characteristics across the country.

[12] Free or reduced price lunch eligibility in Portugal is determined by falling within income brackets that also determine eligibility for other social transfers.

Table 1. Summary Statistics of the Main Analysis Sample

	Observations	Mean / %	Std. Dev.
Outcome:			
Repeat the Grade (%)	766,054	7.73	26.70
Independent Variables of Interest:			
Cumulative Exposure to Students with CE Parents (%)	766,054	18.79	13.83
Contemporaneous Exposure to Students with CE Parents (%)	765,698	19.59	14.34
Mean Cumulative Exposure of Siblings (%)	766,054	18.73	13.68
Cumulative Exposure to Students with Free or Reduced Price Lunch [FRPL] (%)	766,054	41.26	18.58
Cumulative Exposure to Immigrant Students (%)	766,054	4.26	5.22
Cumulative Exposure to Students with CE Parents, in class (%)	766,054	18.33	15.18
Individual Controls:			
Female (%)	766,054	48.61	49.98
Age (in Years)	766,054	9.78	2.58
Free or Reduced Price Lunch [FRPL] (%)	766,054	59.32	49.12
Immigrant (%)	766,054	1.96	13.87
Internet at Home (%)	766,054	51.58	49.98
Number of Siblings	766,054	1.68	0.93
Birth Order	766,054	1.86	0.89
First Born (%)	766,054	38.58	48.68
Difference in Years to First Born	766,054	3.48	4.10

Notes. The sample includes all public school students in mainland Portugal that can be followed since Grade 1, between 2006 and 2016, have at least one sibling enrolled in the public education system in the same academic year, and whose parents have no college education. See Section 2.3 for further details on sample restrictions.

For identification, we exploit variation in peer composition across cohorts within the family. Each academic year, incoming cohorts display variation in student composition.[13] Within the same school, year-on-year changes reflect shocks to school demand due to parental preferences and idiosyncratic changes in neighborhood demographics. Parents start by enrolling their first-born in school. Those with multiple children subsequently enroll younger siblings throughout the following years, usually in the same school as the first-born.[14] Siblings starting school in different academic years will thus be exposed to different levels of cumulative exposure.

Our main identification assumption is that these differences in cumulative exposure between siblings, after controlling for differences across schools and grades within the same academic year, are as good as randomly assigned. Under this conditional independence assumption, the identifying variation comes from siblings enrolled in different grades in the same school or in the same grade but different schools in the same academic year. By controlling for grade-by-year fixed effects, we account for trends in exposure within each grade. School-by-year fixed effects absorb all variation across schools in each year.

3.2. Main Specification

We identify the effect of exposure to students with college-educated (CE) parents by restricting the sample to students without CE parents who have at least one sibling observed in the same academic year. Our fixed effects specification is defined as follows:

[13] The standard deviation in the share of children with CE parents within schools in Portugal is over 8%.

[14] Appendix Figure A.2 shows the distribution of family size among in our full sample. Among our identified 767, 906 families, 50.2 percent of families have more than one child. Among the families with more than one child, 85 percent enroll all siblings in the same public school in first grade.

$$Y_{isgt} = \alpha + \gamma_{gt} + \theta_{st} + \lambda_{ft} + \beta E_{isgt} + \mathbb{X}_{isgt}\delta + \varepsilon_{isgt}. \tag{2}$$

The outcome variable Y_{isgt} represents the outcome of interest, typically grade repetition, for student i, in school s, grade g, and year t. Our main regressor of interest is the cumulative exposure to peers with CE parents (E_{isgt}), as defined in Equation 1. We also include grade-by-year fixed effects (γ_{gt}), school-by-year fixed effects (θ_{st}), and family-by-year fixed effects (λ_{ft}), as according to our main identifying assumption.

To improve precision and to test how sensitive the coefficient of interest (β) is to individual-level controls, we also include a vector of individual-level covariates ($\mathbb{X}$), namely gender, immigration status, eligibility for free or reduced-price lunch (FRPL), and birth order fixed effects.[15] For inference, we cluster the standard errors at the cohort level (school-by-grade-by-year) to account for within-cohort correlation.

Interpretation.— A fundamental concern is how to interpret the estimated β in Equation 2. Our exposure measure captures the potential for children from non-college-educated households to engage with peers from college-educated families. In the data, we have no direct measure of whether these relationships are actually established, so we cannot account for compliance with this plausibly exogenous assignment. As noted by Bifulco et al. (2011), to the extent that social networks form endogenously, we do not aim to establish the causal link of this specific social contagion channel, as these cannot be separately identified from other contextual factors that can change with the variation we exploit.[16] Instead, we interpret our estimates as intention-to-treat effects, acknowledging that these cannot be disentangled from other contextual factors influenced by the variation we exploit. In Section 5 we discuss and provide suggestive evidence for some potential mechanisms.

Crucially, the exposure effect we estimate can be driven by parental education or by other correlated characteristics of highly educated households, such as income. We argue that exposure to households with higher parental education necessarily implies exposure to other characteristics of highly-educated families, influencing the interpretation of our estimate. To gauge how much parental education still stands relative to other household features we report robustness checks that control for cumulative exposure to immigrants and economically disadvantaged students (Section 4.2).

3.3. Identification Challenges

The ability of our empirical strategy to identify the causal parameter of interest faces mainly five concerns: sufficient identifying variation, selection into identification, selection into exposure, within-family spillovers, reflection bias, and exclusion bias. In this section, we discuss each of these challenges

[15] While some of these individual controls vary at the family-level, there may still be variation within-family, for instance, on whether the student is foreign-born or benefits from FRPL, beyond gender and birth order.

[16] As, for instance, Wang (2023), which develops a method to identify the causal effect of forming particular social links.

and present evidence in favor of our identifying assumptions.

Sufficient Variation.— To allay concerns about limited variation, we quantify and characterize residual variation after absorbing all fixed effects in the main specification (Appendix Table A.8, Appendix Figure A.6). Including all fixed effects absorbs 70% of the total variation in cumulative exposure.[17] Importantly, the standard deviation of the residuals of our most stringent specification is of 4p.p.. Thus, while the estimates are driven by small changes in cohort composition, there is sufficient variation in the residuals to produce precise estimates of our parameter of interest (see Section 4.1).

Selection into Identification.— Not all families have variation in peer cumulative exposure across siblings.[18] In our preferred specification, cases with no variation in the residualized measure of cumulative exposure may include twins with identical educational pathways or siblings in schools with no exposure to students with CE parents. Larger families, on the other hand, are more likely to have greater variation in exposure across siblings. Since effects are not identified when there is no within-family-by-year variation, non-random selection of families into the identifying sample may occur. To the extent that the estimated effects may be heterogeneous across groups and selection into identification is substantial, the estimated β may be biased relative to the parameter of interest to our population (Miller et al., 2021). We find that the observations for which we have identifying variation are significantly more likely to repeat the grade, be male, poor, and come from larger families relative to those with no variation (Appendix Table B.9).[19] In light of this evidence, our findings are qualified by the fact that there is some selection into identification. However, observations for which there is non-negligible identifying variation are the large majority (about 90%) of our analysis sample.

Selection into Exposure.— Our main identifying assumption has to be empirically plausible. We conduct informal tests to suggest that omitted variable bias is unlikely in our context. In particular, we examine whether changes in cumulative exposure to peers with CE parents are correlated with observable characteristics when conditioning on our full set of fixed effects. Considering the level of selection on observables as an indication of the level of selection on unobservables, a vanishing correlation between covariates and the regressor of interest under our design lends it credibility.

In Figure 2, we show that there are no relevant differences in cumulative exposure across socioeconomic groups when residualized of the fixed effects in our main specification. We plot absolute stan-

[17] Appendix Table A.8 presents summary statistics of residuals across different models. Appendix Figure A.6 illustrates the distributions of residuals across different models. Removing school-by-year and grade-by-year fixed effects reduces the standard deviation in cumulative exposure by 50%, from 0.14 to 0.07. Adding family-by-year fixed effects absorbs about 70% of the total variation. Moreover, 90% of this residual variation occurs between −0.07 and 0.07 (Appendix Table A.8).

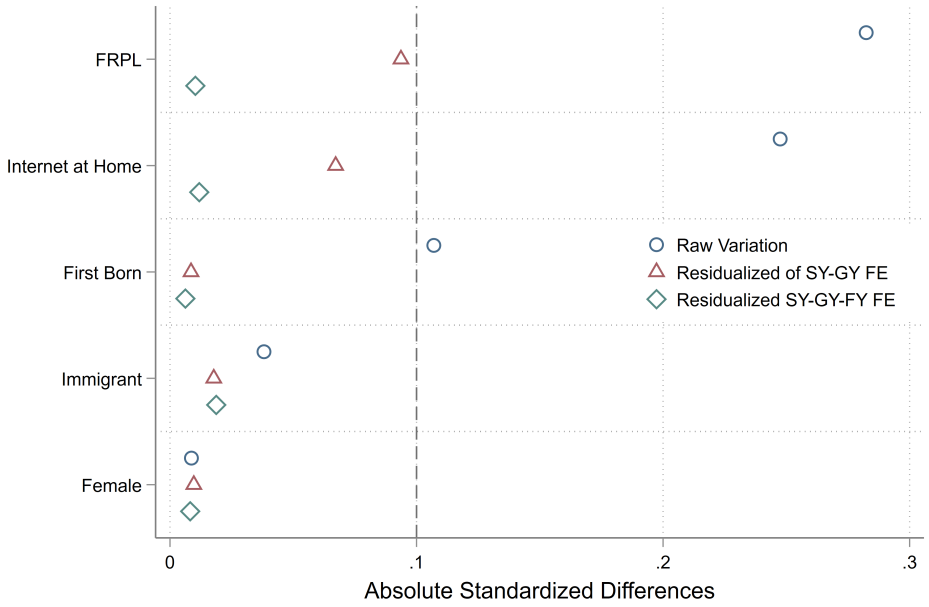
[18] Appendix Figure A.6 shows considerable bunching in the distribution of residuals close to zero, indicating cases where siblings have very similar exposure levels.

[19] In Appendix B we analyze the potential influence of this type of selection. As in Miller et al. (2021), we analyze this potential selection bias by comparing "non-switchers" (families with little to no variation) and "switchers" (families with some variation). Contrary to the cases in Miller et al. (2021), we do not have a binary treatment. However, we use the residuals of our preferred specification to identify other variation close enough to zero. We find that non-switchers are just 11% of the sample, according to our definitions.

dardized differences (Imbens and Wooldridge, 2009) between those with and without a particular demographic characteristic, allowing for different types of comparisons. The dash line indicates a benchmark 0.1 standardized difference as the threshold above which differences across groups are potentially problematic (Austin, 2011). Using all possible variation in the data, we see that there are large differences in exposure based on households' socioeconomic characteristics, such as being eligible to free or reduced-price lunch (FRPL) and having internet at home (circle markers). Figure 2 shows that our design effectively breaks any strong correlations in terms of observables—as measured by the variation residualized of school-by-year, grade-by-year and family-by-year fixed effects (diamond markers). In particular, it breaks the correlation between observables and exposure that would remain if using a within-school, across-cohort design (triangle markers). Alternatively, we also check for balancing on observables by regressing cumulative exposure on a series of students' characteristics, with qualitatively similar diagnostic conclusions (Appendix Table B.10). Finally, we verify whether our main coefficient of interest in Equation 2 is robust to including observable covariates (see Section 4.1).

Importantly, endogeneity may still be a concern if parents strategically allocate siblings to schools with different peer composition based on unobservable ability differences. We carefully address this possibility in Section 4.1.

Figure 2. Differences in Cumulative Exposure Across Groups



Notes. The figure depicts absolute standardized differences between those with and without the specific socioeconomic characteristics in the vertical axis in terms of (i) cumulative exposure to students with college-educated parents (E); (ii) E residualized of grade-by-year and school-by-year fixed effects; and (iii) E residualized of grade-by-year, school-by-year, and family-by-year fixed effects. Standardized differences for each characteristic X are computed as $\left|\frac{\bar{E}_{X=1}-\bar{E}_{X=0}}{\sqrt{(\text{Var}(E_{X=1})+\text{Var}(E_{X=0}))/2}}\right|$, where $\bar{E}_{X=j}$ is the mean exposure for those where $X = j, j \in \{0, 1\}$. The dash line represents a standardized difference of 0.1.

Intrahousehold Spillovers.— Because we rely on comparisons across siblings, spillovers within the

household may contaminate our estimate of interest. Relying on within-family differences, we cannot separately identify the direct effect of peer exposure from the influence that siblings' exposure may have on individual outcomes. Conditional on both these effects going in the same direction and a smaller influence of siblings exposure, we interpret $\hat{\beta}$ as a lower bound of our effect of interest. Appendix C.2 presents a detailed discussion of this concern, and reports evidence in favor of these assumptions. We further discuss these intrahousehold spillovers in Section 4.1.

Reflection Bias.— In the estimation of peer effects, a reflection bias can emerge when treatment and outcomes are measured contemporaneously (Manski, 1993). For instance, each individual's ability can simultaneously cause and be caused by their peers' ability. However, because we rely on exposure to a predetermined variable—parental college education—predictive of, but not contemporaneous peer ability, reflection is not a concern in our setting.

Exclusion Bias.— Recent literature has identified a mechanical downward bias in the estimation of peer effects, even in a context of random group assignment (Caeyers and Fafchamps, 2024; Guryan et al., 2009). In contexts where individuals and the peer group are drawn from the same population, there is an inherent bias as each individual is drawn without replacement. For instance, given a certain distribution of ability in the population, a high (low) ability individual is more likely to being exposed to lower (higher) ability peers, on average. This leads to a negative correlation between individual characteristics and the average characteristic of the individuals in the peer group.

However, with our research design exclusion bias is not a concern. Because we only include in our sample students without CE parents, assignment to groups with different concentrations of peers with CE parents is not affected by the characteristics of the individuals in our sample.

4. The Effect on Grade Repetition

In this section we present our estimates of the effect on educational attainment. We find that higher exposure to peers with college-educated parents decreases the likelihood of grade repetition for children from less educated households. We also show that our findings remain robust across different definitions of exposure and specifications.

4.1. Main Results

Table 2 reports our main results. Across all specifications we keep the same sample of students without CE parents, with at least one sibling who is also enrolled in a public school in the same academic year (see Table 1). As a benchmark, in Column 1, we report the estimated endogenous linear relationship between exposure to students with CE parents and grade repetition (Appendix Figure A.3). The coefficient of −0.09 (s.e. = 0.026) implies that a standard deviation increase in cumulative exposure (13.8 p.p.) is associated with a 1.2 p.p. decrease in grade repetition, or 15.5% over the sample mean.

In Column 2, we present a specification in which we include school-by-year and grade-by-year fixed

Table 2. Grade Repetition and Exposure to Students with College-Educated Parents

	Outcome: Grade Repetition			
	(1)	(2)	(3)	(4)
Exposure to CE	-0.091	-0.124	-0.029	-0.028
	(0.0026)	(0.0045)	(0.0063)	(0.0063)
R-squared	0.002	0.109	0.610	0.611
Observations	766,054	766,054	766,054	766,054
Mean Dep. Var.	.077	.077	.077	.077
School-by-Year FE	No	Yes	Yes	Yes
Grade-by-Year FE	No	Yes	Yes	Yes
Family-by-Year FE	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes. The table reports estimates of β, for alternative specifications of the preferred model presented in Equation 2. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the cumulative exposure to students with CE parents since Grade 1 until the current grade, as defined in Equation 2. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

effects. We find a stronger negative association between our exposure measure and grade repetition, in which a standard deviation increase in cumulative exposure leads to a decrease of about 1.7 p.p. in the probability of repeating the grade. On the one hand, school-by-year fixed effects absorb all across-school variation in grade repetition policies, as well as differences in exposure to students with CE parents, for each academic year. On the other hand, grade-by-year fixed effects take care of differences across grades in each year. In this case, we compare students within the same school and academic year across different grades, or students within the same grade and academic year across different schools. Crucially, parents can still differentially sort into schools on the basis of grade and school composition. Some families without college-educated parents may strategically sort into neighborhoods and schools with stronger exposure to families with CE parents.

In Columns 3 and 4, we show estimates from our preferred specifications. Family fixed effects control for sorting across families, but also underlying cognitive and non-cognitive traits that are common across siblings. With the interaction between family and year, and following Figlio et al. (2023a), we restrict these within-family comparisons to also be within-year, controlling for potential life-cycle changes within the family. We estimate a coefficient of −0.029 (s.e. = 0.0063) (Table 2, Column 3). Reassuringly, the estimate does not not change substantially when including individual controls (Table 2, Column 4). The estimated effect is precisely estimated, statistically significant ($p < 0.01$), and with a relatively nar-

row 95% confidence interval, [−0.041, −0.017]. Appendix Figure E.2 also shows that the non-parametric relationship between our outcome and main independent variable, residualized of school-by-year, grade-by-year and family-by-year fixed effects, is well approximated by a linear fit.

The magnitude of our estimate leads us to conclude that exposure to students with college-educated parents offers some protection against grade repetition for children from less educated households. To understand the economic significance of our effects we can perform a ceteris paribus thought experiment of moving a student across schooling pathways with different concentrations of peers with college-educated parents. Moving from the 10th (2.9%) to the 90th percentile (37.8%) of the distribution of cumulative exposure to students with CE parents would decrease the likelihood of repetition by about 1p.p, or 13% of the sample mean.[20] An effect this size corresponds to about a fifth of the gap in grade repetition between children with and without CE parents (see Section 2.4).

Within-Family Spillovers.— We interpret our estimates as the lower bound of the parameter of interest. With our research design we cannot separately identify the direct effect of peer exposure from the spillover effect of siblings exposure.[21] The effect of higher peer exposure on an individual may also influence their siblings. Empirical evidence shows that siblings influence each others' behavior in educational contexts in multiple ways (e.g. Figlio et al., 2023b; Altmejd et al., 2021). As long as the spillover effect has the same sign and is smaller than the direct effect, comparisons across siblings will underestimate the direct effect of peer exposure on the individual (see Appendix C.2). To provide evidence in support of these conditions we design an alternative specification without family-by-year fixed effects, but in which we separately include individual and average siblings exposure (Equation C.10, Appendix Table C.1). Independently of the child's birth order, the coefficient of siblings' exposure is consistently smaller and the same sign as the coefficient of own exposure on individual outcomes, lending credibility to our interpretation.

Strategic Enrollment.— An important endogeneity concern is whether families strategically enroll their children in different schools based on cohort peer composition. Unobserved differences across siblings may drive strategic enrollment decisions of parents making different human capital investments by child (Becker and Tomes, 1976). For example, parents may enroll higher-achieving children in schools with more college-educated peers due to perceived higher returns on investment, which could bias our estimates upward. Conversely, if parents with an egalitarian approach enroll their lower-achieving children in more exposed schools our estimates could be downward-biased. However, we argue that strategic enrollment is unlikely to be significant in our setting. Priority placement based on geographic catchment areas and having an older sibling in the same school limit the incentives to enroll siblings in different

[20] This is equivalent to moving from an average-sized cohort of 80, with only 2 children with college-educated parents, to a similarly sized cohort with 30 children with CE parents.

[21] In Appendix C.2 we formally present the argument for interpreting the effect as a lower bound.

schools.[22] Moreover, enrolling siblings in different schools can be motivated by reasons other than cohort composition. Notably, school closures—highly prevalent in our period of analysis—force some parents to enroll children in different schools, when reaching the same grade.[23] Finally, while our research design allows for identifying variation to come from siblings enrolled in different schools, school-by-year fixed effects take care of differences in peer composition across schools in the same year. Potential endogeneity would be concerning only if parents enroll children of different unobserved ability in specific school cohorts based on their present or expected peer composition. Therefore, in terms of identification, the possibility of strategic enrollment within the family using this research design does not cost us more than what a within-school, across-cohort strategy would.

Is it College Education?— It could be that other household characteristics are driving the effects. Evidently, families with college-educated parents are different from other families in multiple dimensions other than parental education. In particular, households with highly educated parents are less likely to be poor and foreign-born. To understand whether our estimated effect is driven by these other characteristics, we extend the models presented in Table 2 to include controls of cumulative exposure to immigrant students and recipients of free or reduced price lunch (FRPL). Table 3 report the main coefficients of interest. In our preferred specifications (Columns 3 and 4), controlling for these additional measures of cumulative exposure only slightly attenuates our point estimates, which remain significant ($p < 0.01$). These findings suggest that exposure to students with CE parents still matters for grade repetition beyond differences in exposure to poor and immigrant students. The results in Table 3 also suggest that exposure to FRPL students and immigrants increase the likelihood of grade repetition, although the latter coefficient is imprecisely estimated and not statistically different from zero.

Discussion.— The effect size we uncover is of similar magnitude to estimates on high school completion rates in the US. Bifulco et al. (2011) find that a standard deviation increase in peers with college educated mothers is associated with a 4.7p.p. increase in the likelihood of high school graduation, or 5% over the sample mean. In our case, a standard deviation increase leads to a 0.4p.p. decrease in the likelihood of repetition, also 5% over the sample mean. However, our estimates are not directly comparable for a few reasons. First, the determinants of high school completion may be different than those of yearly grade repetition. Second, we restrict our focus to the sample of students without college-educated parents, rather than all students. Third, we consider peers in the same grade rather than exclusively in the same class. Finally, our research design uses a different source of identifying variation, taking care of unobserved family heterogeneity.

Importantly, we find that the effect using our preferred specification (Table 2, Column 4) is only about a quarter of the effect size uncovered with a specification controlling just for school-by-year and grade-

[22] In our sample, about 80% of the parents enroll their children in the same school, when in the same grade.

[23] Over 32% of public schools in our sample have closed or have consolidated during our period of analysis.

Table 3. Effect on Grade Repetition, Controlling for Other Exposures

	Outcome: Grade Repetition					
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure to CE					-0.0235	-0.0235
					(0.00647)	(0.00647)
Exposure to FRPL	0.0218	0.0171			0.0123	0.0123
	(0.00477)	(0.00477)			(0.00493)	(0.00493)
Exposure to Immigrants			0.0187	0.0177	0.0149	0.0149
			(0.0144)	(0.0144)	(0.0144)	(0.0144)
R-squared	0.610	0.611	0.610	0.611	0.611	0.611
Observations	766,054	766,054	766,054	766,054	766,054	766,054
Mean Dep. Var.	.077	.077	.077	.077	.077	.077
School-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Grade-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Family-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	No	Yes	No	Yes	No	Yes

Notes. The table reports estimates of β, for alternative specifications of the preferred model presented in Equation 2. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressors of interest are the cumulative exposure to students with CE parents, immigrants, and recipients of FRPL, since Grade 1 until the current grade, in definitions analogous to that in Equation 2. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

by-year fixed effects (Table 2, Column 2). On the one hand, each specification uses a different source of identifying variation. Because we restrict differences across cohorts to within-family comparisons, our estimator is relatively more conservative in the type of variation it allows. To the extent that siblings provide a more plausible counterfactual to an individual than just any other peer in adjacent cohorts, our variation is arguably less confounded. On the other hand, the estimates in each specification may include different spillover effects. With our research design, we explicitly acknowledge that the possibility of within-family spillovers allow us to identify a lower bound of our parameter of interest (Appendix C.2). Therefore, we cannot reject that these differences across designs fully stem from within-family spillovers. However, other types of interference may still be present in within-school, across-cohort types of specifications.

4.2. Robustness

To attest the sensitivity of our main results we run a battery of robustness checks. We divide these by sensitivity to alternative definitions of exposure, samples, specifications and clustering of standard errors.

Missing Parental Education.— In the measure described in Equation 1 and used to report results in Table 2, the total number of students in each grade-school cell restricts the count to students with no missing information on parental education.[24] Since this variable is not missing at random, as students with missing parental education are more likely to be poor, more missing information in parental education may overestimate the measure of cumulative exposure for some students. We compute an alternative, more conservative version of our regressor of interest, for which all students with missing information on parental education are treated as not having college-educated parents. In Appendix Table E.3 we report the same specifications as those of our main results. Using this alternative measure of exposure, we uncover quantitatively similar estimates.

Different Definitions of Exposure.— Our main definition of cumulative exposure assumes that exposure in previous academic years carries the same weight as contemporaneous exposure. As in Figlio et al. (2023a), we also expand our definition of cumulative exposure and use a cumulative exposure function which uses a more general specification depending on a decay parameter (κ):

$$E_{isgt}(\kappa) = \frac{\sum_{\underline{t} \leq t' \leq t} \text{Share of Peers with CE Parents}_{isgt'} \times e^{1-\kappa(t-t')}}{\sum_{\underline{t} \leq t' \leq t} e^{1-\kappa(t-t')}}. \tag{3}$$

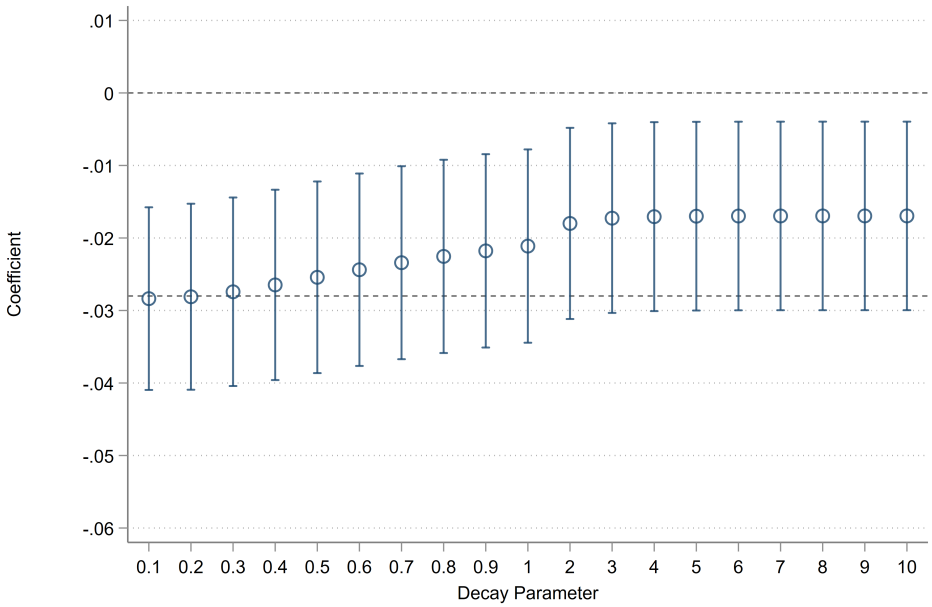
When $\kappa = 0$ the measure in Equation 3 is the same as our main measure of cumulative exposure in Equation 1. However, as κ grows, less weight is given to exposure in previous academic years. Figure 3 depicts the coefficients and the 95% confidence intervals, from regressions with the same specification as Equation 2, using our general measure of exposure, for different values of the decay parameter κ. We conclude that the qualitative interpretation and magnitude of our estimated coefficient is similar across different definitions of exposure. The estimated coefficient monotonically decays with κ, stabilizing in around the same estimated value after $\kappa = 3$. For all values of κ reported, we reject the hypothesis of no effect on grade repetition.

We also investigate whether our results change when using a measure of contemporaneous exposure as our regressor of interest (or equivalently, when $\kappa \to \infty$). Under this alternative specification, we only exploit variation stemming from differences in exposure in a given grade-year cell rather than the average exposure across all grades.[25] Under this type of model, we test what would be the effect if past exposure had no influence in our dependent variable, and all of the effect would be exclusively driven by differences in contemporaneous exposure. Previous literature has mostly relied on this type of contemporaneous variation to estimate spillover effects of parental education (Fruehwirth and Gagete-Miranda, 2019; Bifulco et al., 2011; McEwan, 2003; Bonesronning and Haraldsvik, 2014). Appendix Table E.4 re-

[24] Students with missing information on our measure of parental education account for only 3% of the full sample described in Appendix Table A.1.

[25] Because children are, on average, more exposed to peers with CE parents as they advance throughout grades and academic years, (see Appendix Figure A.1) contemporaneous exposure is relatively larger than cumulative exposure (see Table 1).

Figure 3. Effect on Grade Repetition for Different Definitions of Cumulative Exposure



Notes. The figure depicts estimates of β in Equation 2 in the main text, for different definitions of cumulative exposure to peers with college-educated parents, varying with decay parameter κ in Equation 3. Bars represent 95% confidence intervals. Each value in the x-axis represents a different value of κ. The lower dashed line represents the estimated coefficient using the specification in Table 2, Column 4.

ports our results. We find that the effects are qualitatively similar but relatively smaller in magnitude. We uncover a coefficient of −0.018 (s.e. = 0.0068), or about two-thirds the size of the one in our preferred specification (−0.029, Table 2, Column 5). These findings suggest that the average exposure across grades changes the likelihood of grade repetition not only through contemporaneous exposure.

Finally, we test a definition of exposure which is simply the average number of peers with CE parents across the years, rather than the convex combination of exposure shares:

$$\tilde{E}_{isgt} = \frac{1}{t - \underline{t} + 1} \sum_{\underline{t} \leq t' \leq t} \text{\# Peers with CE Parents}_{isgt'}. \tag{4}$$

It could be that small changes in shares do not translate into actual discrete differences in the number of peers. The exposure definition in Equation 4 allows to estimate the effect of absolute exposure rather than the relative intensity of exposure. Appendix Table E.5 reports our results. We find that the effects are qualitatively similar. With this definition of exposure, we estimate that ten more peers with CE parents in the grade leads to a decrease of half a percentage point in the likelihood of grade repetition. In relative terms, this is similar to the magnitude of our main results, suggesting that our preferred estimates are not driven by the choice of shares in the definition of exposure.

Exposure at the Class Level.— To allay concerns with classroom formation our cumulative exposure variable is measured at the cohort level, not at the class level. Indeed, in Appendix D we show evidence that students are not randomly assigned to classes. Despite this form of potential endogeneity,

we test whether there would be an effect if only classroom peers would influence individual outcomes and the decisions on classroom formation decisions were based on the same observables as we have available. To that end, we run a similar specification to the one in Equation 2, but in which peers are now defined as those within the same class. Under this alternative definition of exposure, we find qualitatively identical results but, as expected, larger in magnitude (Appendix Table D.1, Column 4).

Sensitivity to the Sample.— To allay concerns with our panel of students being unbalanced, we test how each grade contributes to our estimated effect. We start by documenting how the effect changes by iteratively removing observations from each grade from our analysis sample. In these specifications, the identifying variation originates from comparisons between siblings, excluding the observations in which they are at a given grade which is excluded (Appendix Figure E.3). Overall, the effect is consistent across samples.

Alternative Specification.— In Appendix Table E.6 we report results with a different specification than the one in Equation 2, without family-by-year fixed effects. For each individual, we create demeaned versions of the outcome and exposure variables (ΔY, ΔE), where we subtract the mean of each individual's siblings for these variables. We then run the following specification:

$$\Delta Y_{isgt} = \alpha + \gamma_{gt} + \theta_{st} + \beta_2 \Delta E_{isgt} + \mathbb{X}_{isgt}\boldsymbol{\delta} + \varepsilon_{isgt}. \tag{5}$$

In this case, β_2 identifies an effect of having relatively higher exposure than one's siblings, but allowing for comparisons across families. In this analysis, we find qualitatively similar results. We also find that the magnitude of the effects is increasing with birth order (Appendix Table E.6, Columns 2-8).

Alternative Clustering.— Could the level at which standard errors are being clustered make us fail to reject the null hypothesis of no effect of cumulative exposure? In Appendix Table E.7 we consider different levels of clustering and find no significant changes in the 95% confidence intervals of our estimate of interest.

4.3. Heterogeneity

We also investigate whether children in different types of families benefit differentially from higher exposure to peers with college-educated (CE) parents. In particular, we focus on heterogeneity by family size, income and the gender mix of siblings in the household. Moreover, we expand our main analysis to include effects of exposure to peers with CE parents among children of a similar family background.

Family Size.— We test whether peers' influence is stronger among larger families without CE parents. Parental investment per child may be lower among families with more children. The quality of school inputs—such as peer composition—may thus have a stronger effect on the human capital of children of these families. Consistent with this hypothesis we find larger effects among families with three

or more siblings. On the other hand, we find small, insignificant effects among families with just two children (Appendix Table E.8). Importantly, families of different size vary along many dimensions. In particular, children in larger families are more likely to be poor, repeat the grade, or have fewer peers with CE parents. Moreover, given our research design, larger families are more likely to have identifying variation across siblings (Appendix Table B.9).[26] However, we find the distributions of residualized exposure by family size to be overlapping and with an identical standard deviation (Appendix Figure E.4). Therefore, any differences in the estimated coefficients are likely not being driven by differences in identifying variation across family size.

Poverty.— The effect may also differ across families of different socioeconomic status. To study the effect by family income we divide the sample in families for which children benefit from free or reduced price lunch (FRPL) and those that do not. FRPL is our best proxy of poverty and income differences in the data. Table 4 reports the estimated effects, separately by these different groups. We find that the magnitude of the effect among those that do not benefit from FRPL is relatively large (−0.03) and statistically significant ($p < .01$), especially given the relatively lower prevalence of grade repetition in this group (4.7%). On the other hand, the effect among poorer students is smaller (−0.02) and more noisily estimated ($p < .05$). As with family size, the distributions of residualized exposure by FRPL eligibility are identical (Appendix Figure E.5). These results suggest that economic distress—or other characteristics with it associated—reduces the potential benefits of higher exposure to peers with CE parents.

Siblings Gender Mix.— We test for differences in our estimated effect by siblings' gender composition. We divide our sample in families having only-male (25%), only-female (22%), or a mix of male and female children (53%). Appendix Table E.9 reports the results of this analysis. We find that the magnitude of the effect of exposure to peers with CE parents on grade repetition is larger among families with only-male (−0.039) and both male and female children (−0.29). On the other hand, we estimate small (−0.007) and not statistically significant results for families with only girls.

Parental Education.— The main results (Table 2) refer to students whose parents have no college-education, as it pertains to our research question. We replicate our analysis to all children, including those with CE parents, while acknowledging the downward bias introduced by exclusion bias (see discussion in Section 3.3). Appendix Table E.10 reproduces the specifications reported in Table 2 in a sample including all children. We find that the estimates are in all identical to the ones in our main analysis. Indeed, students with CE parents also seem to benefit from greater exposure to peers with similarly educated parents, even though grade repetition is a much rarer event in this subgroup (Appendix Table E.11).

[26] See discussion on selection into identification in Section 3.

Table 4. Heterogeneity by FRPL Status

	Without FRPL			With FRPL		
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure to CE	-0.0770	-0.0263	-0.0308	-0.124	-0.0113	-0.0221
	(0.00591)	(0.00923)	(0.00920)	(0.00706)	(0.0104)	(0.0104)
Std. Coef.	-.053	-.018	-.021	-.055	-.005	-.01
R-squared	0.165	0.657	0.659	0.140	0.625	0.628
Observations	256,395	256,395	256,395	403,690	403,690	403,690
Mean Dep. Var.	.047	.047	.047	.097	.097	.097
School-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Grade-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Family-by-Year FE	No	Yes	Yes	No	Yes	Yes
Individual Controls	No	No	Yes	No	No	Yes

Notes. The table shows regression coefficients analogous to those of Table 2 in the main text. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the cumulative exposure to students with CE parents since they are first enrolled in Grade 1. In Columns 1 through 3, the sample only includes individuals who do not benefit from free or reduced price lunch (FRPL). In Columns 4 through 6, the sample only includes individuals that benefit from FRPL. The standardized coefficient shows the effect of a standard deviation change in the independent variable in terms of standard deviations of the dependent variable. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

5. Mechanisms

This section explores some channels through which greater exposure to peers with college-educated (CE) parents might reduce grade repetition. Our findings in Section 4 are reduced-form estimates of the parameter of interest. Our analysis has not yet addressed the behavioral foundations of this estimated effect. In particular, we consider two potential channels: (i) social contagion and (ii) contextual effects. On the one hand, individuals may adopt the beliefs, behavioral traits and social norms of their peers, either due to a preference for conformity or through direct spillover benefits from exposure to higher-achieving peers (Boucher et al., 2024; Boucher, 2016; Akerlof, 1997). Increased exposure may award children without CE parents more opportunities to engage with peers who have different academic aspirations and behavior. On the other hand, schools might respond differently to higher concentrations of children with CE parents. For example, a higher proportion of relatively advantaged students might influence teaching pace, grade repetition policies, or the grouping of students across classes, with potentially ambiguous effects on individual outcomes. While our research design and data do not allow us to separately identify each causal channel—especially because we cannot observe students' social networks—we do have rich information on school policies, such as leniency towards students eligible to repeat or children's academic performance on different school subjects. Using this information, we provide suggestive evidence

that may support or challenge each of these mechanisms.

5.1. Learning

It could be that the reduced risk of grade repetition is fully driven by contextual effects and does not stem from actual learning gains. We start by examining whether greater exposure to peers leads to better academic achievement upstream of grade repetition. We restrict the sample to observations in middle schools—between fifth and ninth grade—for which we have information on performance in each school subject. In this analysis, we thus rely on an unbalanced panel of siblings without CE parents who are enrolled in a public middle school in the same academic year. Thus, an important caveat is that the sample used in this analysis is different and considerably smaller than the one in Section 4.[27]

A first important channel is students' overall academic performance in school. In our setting, we do not have—for every grade—standardized exam results which can be compared across schools in each year. Instead, we use a measure of achievement at the school level: students' grade point average (GPA) in mandatory core subjects.[28]

A second relate channel is whether students are eligible to repeat. To become eligible students must either fail at three or more school subjects or fail at both Math and Portuguese Language in fourth, sixth or ninth grade (see Section 2.4). Each year, over a third (35%) of students in our sample fail to attain a passing grade in Math. On the other hand, 16% fail Portuguese Language. Across mandatory core subjects, each individual in each grade fails on average in 12% of these courses.[29] On average, 15.7% of the students are eligible to repeat the grade, each year. However, only slightly more than half of these (8.9%) actually repeat the grade.

Table 5 reports the main results of our middle school analysis. In each column, we use the same type of specification as in Equation 2. We find that siblings exposed to a higher share of peers with CE parents experience a statistically significant GPA increase (p-value < 0.01), reflecting a positive effect on overall academic performance. Interpreting this effect in terms of the standardized beta coefficient, we conclude that a standard deviation (12p.p.) increase in cumulative exposure leads to a 2.6 percent of a standard deviation (0.59) increase in GPA (Table 5, Column 2).

Our analysis also reveals that siblings exposed to greater proportions of students with CE parents tend to perform better across multiple subjects, failing at a smaller proportion of subjects (Table 5, Column 6). Crucially, two of these subjects at which more exposed students fail less are Language and Math.

[27] Appendix Table A.6 reports summary statistics of this restricted sample.

[28] In each subject, students are awarded a grade between 1 and 5 by their teachers. GPA averages all these grades in different subjects for each student. The main limitation of this outcome variable is that it depends on teacher-specific and school-specific policies. However, in our main specifications, we control for school-by-year and grade-by-year fixed effects, allaying some of these concerns.

[29] In grades 5 and 6, mandatory core subjects are: Language, Math, Physical Education, Natural Sciences, English, History and Geography, Drawing and Technological Education. Between grades 7 and 9 mandatory core subjects are: Language, Math, Physical Education, Natural Sciences, English, Second Foreign Language, History, Geography, Physics and Chemistry, Drawing, and Technological Education.

Table 5. Effects on Academic Performance in Middle School

Outcomes:	Repeat	GPA (1-5)	Fail Language	Fail Math	Fail Language and Math	Prop. of Failed Subjects	Eligible to Repeat
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Exposure to CE	-0.0193	0.128	-0.0719	-0.0562	-0.0506	-0.0542	-0.0608
	(0.0210)	(0.0361)	(0.0275)	(0.0334)	(0.0234)	(0.0126)	(0.0261)
Std. Coef.	-.008	.026	-.023	-.014	-.019	-.034	-.02
R-squared	0.612	0.769	0.620	0.665	0.613	0.706	0.639
Observations	134,548	134,548	134,548	134,548	134,548	134,548	134,548
Mean Dep. Var.	.089	3.349	.163	.345	.111	.128	.157
School-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Family-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. The table reports estimates of β, for alternative specifications of the preferred model presented in Equation 2. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 5 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable varies, as indicated in the header of each column in the table. GPA (1-5) is a continuous measure of the grade point average of students in a set of core subjects in school. Fail Language, Fail Math, and Fail Both are dummy variables that indicate whether the student failed (i.e., had a grade lower than 2, in a scale from 1 to 5) in each of the subjects, respectively, and at both subjects. Proportion of Failed Subjects measures the proportion of core subjects at which the student had a grade lower than 2. Eligible to Repeat indicates whether the students fulfills the criteria to fail the grade. The regressor of interest is the cumulative exposure to students with CE parents, since Grade 1 until the current grade, as defined in Equation 2. The standardized coefficient shows the effect of a standard deviation change in the independent variable in terms of standard deviations of the dependent variable. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

A standard deviation increase in exposure to students with CE parent (12p.p.) leads to a 5% decrease in the likelihood of failing at Portuguese Language (Column 3). The probability of failing at Math decreases by half a percentage point, but the coefficient is noisily estimated (Column 4). Moreover, students are also less likely to concurrently repeat at both Language and Math in the same academic year (Column 5), an effect that is strongly statistically significant ($p < .01$). Finally, the likelihood to of being eligible to repeat a grade also decreases: a standard deviation increase in exposure to students with CE parents leads to a decrease of 4.5% decrease over the mean of this variable (Column 6). In this restricted sample, for which we have no power to detect a negative effect on actual grade repetition (Column 1, $p = .35$), moving a student from the 10th (4%) to the 90th percentile (34%) in cumulative exposure to peers with CE parents decreases the likelihood of being eligible to repeat by 1.8 percentage points.

Discussion.— Overall, these findings lead us to reject the hypothesis that greater exposure to CE parents has no effect on the school performance of relatively less advantaged peers. We argue that the magnitude of these effects is relatively small but in line with other estimates in the empirical literature on peer effects. In high school, Bifulco et al. (2011) do not find any significant association of exposure to highly educated parents on GPA. Among Chilean eight graders, a standard deviation increase in

the classroom mean of mothers' education leads to a 0.27 standard deviations improvement in reading achievement (McEwan, 2003). In Norway, a standard deviation increase in exposure to less educated parents leads to a 0.02 standard deviation decrease in achievement among fifth graders (Bonesronning and Haraldsvik, 2014). However, the evidence is not directly comparable, given differences in sample definitions and research designs. Conditional on a similar empirical strategy, our effect size is larger than an estimated positive coefficient of exposure to foreign-born children among native US students (Figlio et al., 2023a): 1.2% of a standard deviation in math test scores vs. 2.6% of a standard deviation in GPA in our setting.

5.2. Grade Repetition Policies

In Section 5.1 we show that part of our results are driven by learning gains. In Section 5.?? we show suggestive evidence that effects also differ by how students are grouped into classes. In this section we study whether part of the reduction on grade repetition is driven by changes in retention policies at the school level.

As discussed in Section 2, and reported in Appendix Table E.1, we show that schools have considerable leeway in deciding on who to repeat, conditional on a student being eligible to being retained in the same grade. Overall, only 60% of eligible students actually repeat the grade. It could be that enrolling more students with CE parents makes schools become more lenient with respect to grade repetition. To the extent that more educated parents put pressure on schools not to repeat their own children when eligible, schools may become more lenient in their repetition policies to all enrolled students.

To test this hypothesis, we define a cohort-level measure of leniency towards children without college-educated parents for every school in the sample. We define leniency as the proportion of eligible students without CE parents in cohort sgt that are eligible, but do not repeat the grade:

$$L_{sgt} = \frac{\sum_{i \in sgt} \text{NCE}_i \times (1 - \text{Repeat}_i)}{\sum_{i \in sgt} \text{NCE}_i \times \text{Eligible}_i}, \tag{6}$$

where NCE_i indicates whether student i in cohort sgt does not have college-educated parents. As $L_{sgt} \to 1$, the higher the proportion of children without college educated parents that, being eligible, do not repeat the grade. The mean value of L_{sgt} in our sample is .445, implying that, on average, 44.5% of these eligible students do not repeat the grade. Taking our measure of leniency as an outcome, we then run variations of the following specification at the cohort-level:

$$L_{sgt} = \alpha + \gamma_{gt} + \theta_{sg} + \eta \text{Prop. CE}_{sgt} + \delta \text{Eligible w/o CE}_{sgt} + \epsilon_{sgt}, \tag{7}$$

where Prop. CE_{sgt} is the proportion of students with college-educated parents in the cohort, and Eligible w/o CE_{sgt} is the number of students without CE parents in the cohort who are eligible to repeat the grade. Depending on the specification, we also include grade-by-year (γ_{gt}) and school-by-grade fixed

effects (θ_{sg}).

Table 6. Leniency and Proportion of Students with CE Parents

	Outcome: Grade Repetition Leniency				
	(1)	(2)	(3)	(4)	(5)
Prop. with CE Parents	0.210	0.263	0.218	0.271	0.194
	(0.0171)	(0.0371)	(0.0171)	(0.0372)	(0.0379)
Eligible to Repeat w/o CE Parents			0.004	0.003	-0.000
			(0.0002)	(0.0003)	(0.0003)
Std. Coef.	.088	.111	.092	.114	.082
R-squared	0.008	0.508	0.020	0.510	0.562
Observations	21,309	21,309	21,309	21,309	21,309
Mean Dep. Var.	.445	.445	.445	.445	.445
School-by-Year FE	No	Yes	No	Yes	Yes
School-by-Grade FE	No	No	No	No	Yes

Notes. The table reports estimates of η and δ, for alternative specifications of the preferred model presented in Equation 7. The sample is an unbalanced panel of middle schools in mainland Portugal, between Grades 5 and 9. The dependent variable is leniency in grade repetition in the end of a given grade-year, as defined in Equation 6. The regressor of interest is the contemporaneous share of students with CE parents. The standardized coefficient shows the effect of a standard deviation change in the independent variable in terms of standard deviations of the dependent variable. Robust standard errors are presented in parentheses.

Table 6 reports the associations between leniency and the share of CE parents. A standard deviation increase in the proportion of students with CE parents is associated with a 8.8% standard deviation increase in our measure of grade repetition leniency (Column 1). In Column 2 we control for school-by-year fixed effects to absorb any differences across schools in the same academic year. Thus, in this specification we only use as variation differences in the concentration of children with CE parents within-school, across-grades. In Columns 3 through 5, we further control for the number of eligible students without CE parents to allay concerns with the effect being mostly driven by the concentration of this type of students in the cohort. Across all specifications, schools are more lenient when grade repetition is higher. Overall, these findings suggest that we cannot rule out schools becoming more lenient towards grade repetition, beyond the actual learning gains of more exposed students.

6. Conclusion

We use administrative data from Portugal to study the effect of exposure to children from more educated households on the school outcomes of children whose parents have not completed college. Portugal has one the highest rates of students repeating a grade among OECD countries, with children who do not have parents with a college degree being six times more likely to repeat a grade than those with at least one college-educated parent. The panel structure of the data and family identifiers allow

us to deal with the endogenous sorting of families across schools, as well as recover the full history of exposure across years for each individual. For identification, we compare the outcomes of siblings who are exposed to different levels of exposure to students with college-educated parents which are not explained by differences in the school, grade, or academic year in which they are enrolled. Therefore, our paper goes beyond most of the extant literature, which typically just relies on comparisons across cohorts within schools.

The main finding is that exposure to students with college-educated is a significant protection against grade repetition for children from less educated households. Importantly, we find that simply relying on variation within school across cohorts would make us largely overestimate the effect of exposure on grade repetition. Although being more stringent, our preferred identification strategy still uncovers an effect that is economically significant. The estimated effect size suggests that moving a student from the 10th to the 90th percentile in the distribution of cumulative exposure to students with CE parents corresponds to about a fifth of the large gap in grade repetition between children with and without CE parents. Furthermore, exposure to students with CE parents matters for educational attainment beyond differences in exposure to poor and immigrant students.

We find evidence that our cumulative exposure measure affects grade repetition through greater student performance in school across all subjects. A standard deviation increase in exposure to children with CE parents leads to 2.6 percent of a standard deviation increase in GPA, and 3.4 percent of a standard deviation decrease in the proportion of failed subjects. However, we cannot exclude the importance of contextual effects that go beyond an improvement in actual student learning. We find suggestive evidence that when schools are faced with higher concentrations of students with college-educated parent they are also less likely to fail children without CE parents, conditional on being eligible to repeating the grade.

References

Akerlof, G. A. (1997): "Social Distance and Social Decisions," *Econometrica*, 65, 1005.

Allen, R., S. Burgess, R. Davidson, and F. Windmeijer (2015): "More reliable inference for the dissimilarity index of segregation," *Econometrics Journal*, 18, 40–66.

Altmejd, A., A. Barrios-Fernández, M. Drlje, J. Goodman, M. Hurwitz, D. Kovac, C. Mulhern, C. Neilson, and J. Smith (2021): "O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries," *Quarterly Journal of Economics*, 136, 1831–1886.

Ammermueller, A. and J. S. Pischke (2009): "Peer effects in european primary schools: Evidence from the progress in international reading literacy study," .

Austin, P. C. (2011): "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, 46, 399–424.

Barrios-Fernández, A. (2023): "Peer Effects in Education," *Oxford Research Encyclopedia of Economics*

and Finance.

Becker, G. S. and N. Tomes (1976): "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy*, 84, S143–S162.

Bertoni, M., G. Brunello, and L. Cappellari (2020): "Who benefits from privileged peers? Evidence from siblings in schools," *Journal of Applied Econometrics*, 35, 893–916.

Bianchi, S. M. and J. Robinson (1997): "What Did You Do Today? Children's Use of Time, Family Composition, and the Acquisition of Social Capital," *Journal of Marriage and the Family*, 59, 332.

Bifulco, R., J. M. Fletcher, and S. L. Ross (2011): "The effect of classmate characteristics on post-secondary outcomes: Evidence from the add health," *American Economic Journal: Economic Policy*, 3, 25–53.

Bonesronning, H. and M. Haraldsvik (2014): "Peer effects on student achievement: Does the education level of your classmates parents matter," *Working paper*.

Boneva, T. and C. Rauh (2018): "Parental beliefs about returns to educational investments-The Later the better?" *Journal of the European Economic Association*, 16, 1669–1711.

Borghesan, E., H. Reis, and P. E. Todd (2022): "Learning Through Repetition? A Dynamic Evaluation of Grade Retention in Portugal," .

Boucher, V. (2016): "Conformism and self-selection in social networks," *Journal of Public Economics*, 136, 30–44.

Boucher, V., M. Rendall, P. Ushchev, and Y. Zenou (2024): "Towards a General Theory of Peer Effects," *Econometrica*, 92, 543–565.

Caeyers, B. and M. Fafchamps (2024): "Exclusion Bias in the Estimation of Peer Effects," *Journal of Human Resources*.

Cascio, E. U. and E. G. Lewis (2012): "Cracks in the melting pot: Immigration, school choice, and segregation," *American Economic Journal: Economic Policy*, 4, 91–117.

Cattan, S., K. G. Salvanes, and E. Tominey (2023): "First generation elite : the role of school networks," *Working Paper*.

Cools, A., R. Fernández, and E. Patacchini (2019): "Girls, Boys, and High Achievers," *NBER Working Papers No. 25763*.

Epple, D. and R. E. Romano (2011): "Peer effects in education: A survey of the theory and evidence," in *Handbook of Social Economics*, vol. 1, 1053–1163.

Figlio, D., P. Giuliano, R. Marchingiglio, U. Ozek, and P. Sapienza (2023a): "Diversity in Schools: Immigrants and the Educational Performance of U.S. Born Students," *Review of Economic Studies*, forthcomin.

Figlio, D. and U. Özek (2020): "An extra year to learn English? Early grade retention and the human capital development of English learners," *Journal of Public Economics*, 186.

Figlio, D. N., K. Karbownik, and U. Özek (2023b): "Sibling Spillovers May Enhance the Efficacy of Targeted School Policies," *NBER Working Paper 31406*.

Fruehwirth, J. C. and J. Gagete-Miranda (2019): "Your peers' parents: Spillovers from parental education," *Economics of Education Review*, 73, 101910.

Guryan, J., E. Hurst, and M. Kearney (2008): "Parental education and parental time with children," in *Journal of Economic Perspectives*, vol. 22, 23–46.

Guryan, J., K. Kroft, and M. J. Notowidigdo (2009): "Peer effects in the workplace: Evidence from random groupings in professional golf tournaments," *American Economic Journal: Applied Economics*, 1, 34–68.

Hill, C. R. and F. P. Stafford (1974): "Allocation of Time to Preschool Children and Educational Opportunity," *The Journal of Human Resources*, 9, 323.

Hoxby, C. (2000): "Peer Effects in the Classroom: Learning from Gender and Race Variation," *NBER Working Paper 7867*.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47, 5–86.

Jacob, B. A. and L. Lefgren (2004): "Remedial education and student achievement: A regression-discontinuity analysis," *Review of Economics and Statistics*, 86, 226–244.

——— (2009): "The Effect of Grade Retention on High School Completion," *American Economic Journal: Applied Economics*, 1, 33–58.

Lavy, V. and A. Schlosser (2011): "Mechanisms and Impacts of Gender Peer Effects at School," *American Economic Journal: Applied Economics*, 3, 1–33.

Liebowitz, D., P. González, E. Hooge, and G. Lima (2018): *OECD Reviews of School Resources: Portugal 2018*, OECD Reviews of School Resources, Paris: OECD Publishing.

Manacorda, M. (2012): "The Cost of Grade Retention," *The Review of Economics and Statistics*, 94, 596–606.

Manski, C. F. (1993): "Identification of endogenous social effects the reflection problem," *Review of Economic Studies*, 60, 531–542.

McEwan, P. J. (2003): "Peer effects on student achievement: Evidence from Chile," *Economics of Education Review*, 22, 131–141.

Miller, D. L., N. Shenhav, and M. Grosz (2021): "Selection into Identification in Fixed Effects Models, with Application to Head Start," *Journal of Human Resources*, 0520–10930R1.

Müller, M. W. (2023): "Intergenerational Transmission of Education: Internalized Aspirations versus Parents Pressure," *Job Market Paper*.

Nechyba, T. J. (2006): "Income and Peer Quality Sorting in Public and Private Schools," in *Handbook of the Economics of Education*, Elsevier, vol. 2, 1327–1368.

OECD (2020): "How students progress through schooling," in *PISA 2018 Results: Effective Policies, Successful Schools*, Paris: OECD Publishing, vol. V of *PISA*, chap. 2.

——— (2022): *Review of Inclusive Education in Portugal*, Paris: OECD Publishing.

Owens, A., S. F. Reardon, and C. Jencks (2016): "Income Segregation Between Schools and School Districts," *American Educational Research Journal*, 53, 1159–1197.

Sacerdote, B. (2011): *Peer Effects in Education: How might they work, how big are they and how much do we know Thus Far?*, vol. 3.

Schwerdt, G., M. R. West, and M. A. Winters (2017): "The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida," *Journal of Public Economics*, 152, 154–169.

ter Meulen, S. (2023): "Long-Term Effects of Grade Retention," *CESifo Working Paper No. 10212*.

Wang, Z. (2023): "The linking effect: causal identification and estimation of the effect of peer relationship," *Job Market Paper*.

APPENDICES

A. The Data

The data for this paper relies on a combination of multiple administrative data files. Below we describe all relevant information about the cleaning and coding of the variables used in the analysis.

A.1. Data Sources

Enrollment Data: The enrollment data is organized at the student-by-year level. Each record contains a unique student and family identifier and information on students' birth date, gender, country of origin, address, parent's education and employment status, free and reduced-price lunch eligibility, access to computer and Internet at home, school, class, curriculum, and grade. The student identifier allows to track individuals throughout classes, grades and schools across years, which enables us to collect information about their educational career such as grade repetition, our main outcome of interest. Attrition in the data may occur for a few reasons: If the student moves abroad, drops from the education system altogether, dies, or the matching algorithm is unable to correctly assign the unique identifier to new instance of the same student in the system. However, in following students, attrition rates for different cohorts are relatively limited (below 10% for follow-up periods of ten years). Enrollment data comes from MISI-PUB, MISI-PRIV and INQ-PRIV, administrative datasets with information on every student enrolled in public, publicly-funded private and private schools in mainland Portugal. Information on socioeconomic characteristics of students enrolled in private schools, however, is substantially more limited. The data is available between the academic years of 2006-2007 and 2017-2018.

School Outcomes: The data on school outcomes is organized at the student-by-subject-by-year level. For each subject in lower and upper secondary education, the grade (from 1 to 5) of the student in the subject, awarded by teachers is recorded. From these records we construct students' GPA in each year for which data is available, as well as information on the subjects failed and eligibility for grade repetition.

Data Access: Access to the data is restricted. All data used in this paper is hosted in a server at a safe center in Nova School of Business and Economics, located in Lisbon, Portugal. As of now, accessing this information requires the researcher to be physically present in the safe center.

A.2. Additional Descriptive Statistics

Table A.1. Summary Statistics of the Full Dataset before Sample Restrictions

	Observations	Mean / %	Std. Dev.
Outcome:			
Repeat the Grade (%)	5,272,709	5.31	22.43
Independent Variables of Interest:			
Cumulative Exposure to Students with CE Parents (%)	5,272,709	20.99	16.28
Contemporaneous Exposure to Students with CE Parents (%)	5,050,150	22.46	16.42
Cumulative Exposure to Students with Free or Reduced Price Lunch [FRPL] (%)	5,272,709	36.93	19.09
Cumulative Exposure to Immigrant Students (%)	5,272,709	4.49	5.54
College-Educated Parents (%)	5,053,059	22.34	41.65
Individual Controls:			
Female (%)	5,272,709	48.52	49.98
Age (in Years)	5,272,709	9.28	2.47
Free or Reduced Price Lunch [FRPL] (%)	5,272,709	40.04	49.00
Immigrant	5,272,691	3.29	17.83
Internet at Home (%)	5,272,709	53.17	49.90
Number of Siblings	5,217,283	0.88	0.91
Birth Order	5,217,283	1.50	0.73

Notes. The table preseents information analogous to the one presented in Table 1, in the main text. The sample includes all public school students in Portugal, enrolled in regular curriculum offers, between the years of 2006 and 2016.

Table A.2. Summary Statistics of the Dataset Restricted to Children with Siblings

	Observations	Mean / %	Std. Dev.
Outcome:			
Repeat the Grade (%)	3,230,989	5.19	22.19
Independent Variables of Interest:			
Cumulative Exposure to Students with CE Parents (%)	3,230,989	20.63	16.31
Contemporaneous Exposure to Students with CE Parents (%)	3,112,981	21.96	16.47
Cumulative Exposure to Students with Free or Reduced Price Lunch [FRPL] (%)	3,230,989	37.00	19.24
Cumulative Exposure to Immigrant Students (%)	3,230,989	4.22	5.28
College-Educated Parents (%)	3,114,825	23.70	42.52
Individual Controls:			
Female (%)	3,230,989	48.31	49.97
Age (in Years)	3,230,989	9.35	2.48
Free or Reduced Price Lunch [FRPL] (%)	3,230,989	41.39	49.25
Immigrant	3,230,979	1.98	13.91
Internet at Home (%)	3,230,989	54.44	49.80
Number of Siblings	3,230,989	1.42	0.75
Birth Order	3,230,989	1.81	0.78

Notes. The table preseents information analogous to the one presented in Table 1, in the main text. The sample includes all public school students in Portugal, enrolled in regular curriculum offers, between the years of 2006 and 2016, with at least one sibling observed at least once throughout the period of analysis.

Table A.3. Comparison between the Full Dataset and the Data Restricted to Children with Siblings

	Siblings Sample			Full Sample		
	Observations	Mean / %	Std. Dev.	Observations	Mean / %	Std. Dev.
Outcome:						
Repeat the Grade (%)	3,230,989	5.19	22.19	5,272,709	5.31	22.43
Independent Variables of Interest:						
Cumulative Exposure to Students with CE Parents (%)	3,230,989	20.63	16.31	5,272,709	20.99	16.28
Contemporaneous Exposure to Students with CE Parents (%)	3,112,981	21.96	16.47	5,050,150	22.46	16.42
Cumulative Exposure to Students with Free or Reduced Price Lunch [FRPL] (%)	3,230,989	37.00	19.24	5,272,709	36.93	19.09
Cumulative Exposure to Immigrant Students (%)	3,230,989	4.22	5.28	5,272,709	4.49	5.54
College-Educated Parents (%)	3,114,825	23.70	42.52	5,053,059	22.34	41.65
Individual Controls:						
Female (%)	3,230,989	48.31	49.97	5,272,709	48.52	49.98
Age (in Years)	3,230,989	9.35	2.48	5,272,709	9.28	2.47
Free or Reduced Price Lunch [FRPL] (%)	3,230,989	41.39	49.25	5,272,709	40.04	49.00
Immigrant	3,230,979	1.98	13.91	5,272,691	3.29	17.83
Internet at Home (%)	3,230,989	54.44	49.80	5,272,709	53.17	49.90
Number of Siblings	3,230,989	1.42	0.75	5,217,283	0.88	0.91
Birth Order	3,230,989	1.81	0.78	5,217,283	1.50	0.73

Notes. The table compares the full sample, as described in Table A.1, with the sample of children with at least one siblings, as described in Table A.2.

Table A.4. Summary Statistics of the Dataset Restricted to Children with Siblings in School in the Same Year

	Observations	Mean / %	Std. Dev.
Outcome:			
Repeat the Grade (%)	1,244,527	6.13	23.98
Independent Variables of Interest:			
Cumulative Exposure to Students with CE Parents (%)	1,244,527	21.85	16.63
Exposure to Students with CE Parents (%)	1,198,131	23.31	16.76
Cumulative Exposure to Students with Free or Reduced Price Lunch [FRPL] (%)	1,244,527	38.67	18.47
Exposure to Immigrant Students	1,244,136	4.35	5.45
College-Educated Parents (%)	1,198,576	27.02	44.41
Individual Controls:			
Female (%)	1,244,527	48.34	49.97
Age (in Years)	1,244,527	9.70	2.53
Free or Reduced Price Lunch [FRPL] (%)	1,244,527	46.15	49.85
Immigrant	1,244,527	1.82	13.35
Internet at Home (%)	1,244,527	56.25	49.61
Number of Siblings	1,244,527	1.73	1.01
Birth Order	1,244,527	1.95	0.95

Notes. The table preseents information analogous to the one presented in Table 1, in the main text. The sample includes all public school students in Portugal, enrolled in regular curriculum offers, between the years of 2006 and 2016, with at least one sibling observed at least once in the same academic year.

Table A.5. Comparison between the Siblings Sample with the Sample of Siblings in the Same Academic Year

	Siblings in the Same Academic Year Sample			Siblings Sample		
	Observations	Mean / %	Std. Dev.	Observations	Mean / %	Std. Dev.
Outcome:						
Repeat the Grade (%)	1,174,502	6.12	23.96	3,230,989	5.19	22.19
Independent Variables of Interest:						
Cumulative Exposure to Students with CE Parents (%)	1,174,502	22.09	16.70	3,230,989	20.63	16.31
Contemporaneous Exposure to Students with CE Parents (%)	1,131,462	23.54	16.83	3,112,981	21.96	16.47
Cumulative Exposure to Students with Free or Reduced Price Lunch [FRPL] (%)	1,174,502	38.62	18.42	3,230,989	37.00	19.24
Cumulative Exposure to Immigrant Students (%)	1,174,502	4.20	5.04	3,230,989	4.22	5.28
College-Educated Parents (%)	1,131,897	27.87	44.84	3,114,825	23.70	42.52
Individual Controls:						
Female (%)	1,174,502	48.34	49.97	3,230,989	48.31	49.97
Age (in Years)	1,174,502	9.72	2.54	3,230,989	9.35	2.48
Free or Reduced Price Lunch [FRPL] (%)	1,174,502	46.26	49.86	3,230,989	41.39	49.25
Immigrant	1,174,502	1.81	13.33	3,230,979	1.98	13.91
Internet at Home (%)	1,174,502	56.76	49.54	3,230,989	54.44	49.80
Number of Siblings	1,174,502	1.61	0.89	3,230,989	1.42	0.75
Birth Order	1,174,502	1.82	0.86	3,230,989	1.81	0.78

Notes. The table compares the sample of siblings, as described in Table A.2, with the sample of children with at least one other sibling in the same academic year, as described in Table A.4.

Table A.6. Summary Statistics of Middle School Sample

	Mean / %	Std. Dev.
Outcomes:		
Repeat the Grade (%)	8.85	28.41
GPA (1-5)	3.35	0.59
Fail Math (%)	34.55	47.55
Fail Language (%)	16.25	36.89
Faild Language and Math	11.09	31.41
Failed Subjects (%)	12.79	18.99
Eligible to Repeat (%)	15.65	36.33
Independent Variables of Interest:		
Cumulative Exposure to Students with CE Parents (%)	17.77	12.00
Contemporaneous Exposure to Students with CE Parents (%)	19.41	12.77
Cumulative Exposure to Students with Free or Reduced Price Lunch [FRPL] (%)	43.09	14.26
Cumulative Exposure to Immigrant Students (%)	4.64	4.82
Individual Controls:		
Female (%)	49.67	50.00
Age (in Years)	12.28	1.59
Free or Reduced Price Lunch [FRPL] (%)	65.53	47.53
First Generation Immigrant	2.15	14.52
Internet at Home (%)	58.39	49.29
Number of Siblings	1.76	0.95
Birth Order	1.85	0.88
Observations	134,548	

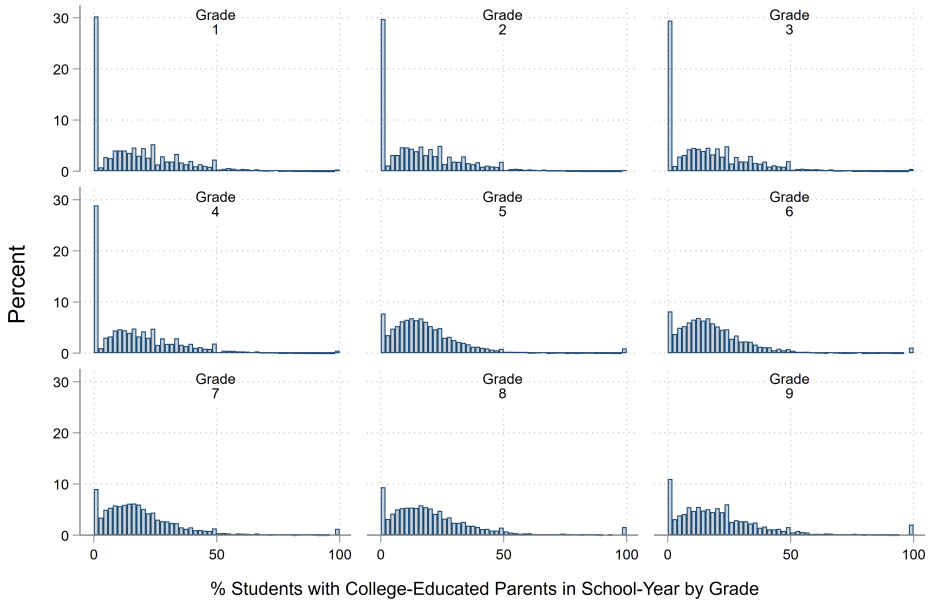
Notes. The sample includes all public school students in mainland Portugal that can be followed since Grade 1, between 2006 and 2016, and have at least one sibling enrolled in a public middle school (between grades 5 and 9) in the same academic year.

Table A.7. Summary Statistics at the Student Level

	Analysis Sample			Siblings Sample			Full Sample		
	Observations	Mean / %	Std. Dev.	Observations	Mean / %	Std. Dev.	Observations	Mean / %	Std. Dev.
Total Nr. of Repetitions	200,457	0.40	0.79	574,650	0.30	0.70	957,555	0.29	0.69
Ever Repeat (%)	200,457	26.59	44.18	574,650	19.83	39.87	957,555	19.72	39.79
Contemporaneous Exposure to Students with CE Parents (%)	200,428	20.28	13.72	550,790	23.71	16.07	913,407	24.22	16.03
Average Class Size	200,457	19.74	4.88	574,650	20.20	4.86	957,555	20.30	4.87
Average Cohort Size	200,457	64.22	38.75	574,650	66.87	39.94	957,555	66.42	39.77
College-Educated Parents (%)	200,457	0.36	4.33	550,868	26.02	43.57	913,541	24.58	42.74
Female (%)	200,457	48.52	49.98	574,650	48.34	49.97	957,555	48.54	49.98
Internet at Home (%)	200,457	50.57	42.19	574,650	55.71	42.06	957,555	54.55	42.41
Immigrant (%)	200,457	1.87	13.54	574,648	3.25	17.74	957,549	3.96	19.51
Free or Reduced Price Lunch [FRPL] (%)	200,457	54.72	40.37	574,650	40.35	41.05	957,555	38.65	40.85
Number of Siblings	200,457	1.59	0.88	558,856	1.41	0.75	941,761	0.84	0.91
Have Twins (%)	200,457	7.51	26.36	558,856	4.18	20.02	941,761	2.48	15.56
Birth Order	200,457	1.85	0.88	558,856	1.85	0.80	941,761	1.51	0.75
Difference in Years to First Born	200,457	3.71	4.33	558,856	4.50	4.67	558,856	4.50	4.67
First Born (%)	200,457	39.00	48.78	574,650	33.80	47.30	957,555	60.27	48.93

Notes. The table compares student-level statistics for different samples: the sample used in the main analysis; the sample of all public education students with at least one sibling; and the population of public school students in Portugal, enrolled in regular curriculum offers, between the years of 2006 and 2016.

Figure A.1. Distributions of the Proportion of Students with CE Parents by Grade



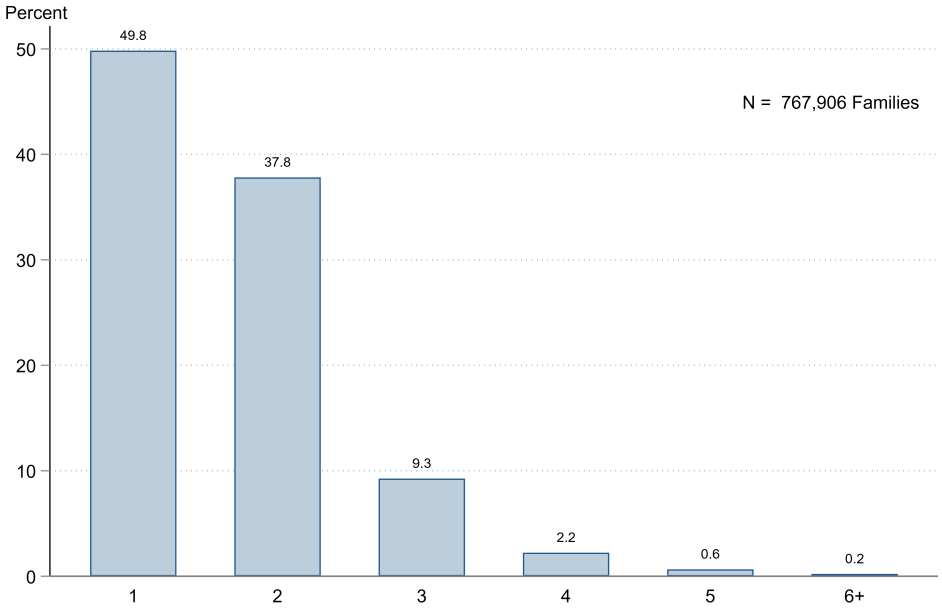
Notes. The figure depicts the distribution of the percentage of students with college-educated parents in each school-by-year cell, for each school grade.

Table A.8. Variation in Cumulative Exposure after Removing Each Set of Fixed Effects

	Mean	Std. Dev.	Minimum	P5	Median	P95	Maximum
Raw Variation	0.19	0.14	0.00	0.00	0.16	0.45	1.00
Excluding SY + GY FE	0.00	0.07	-0.53	-0.11	-0.00	0.12	0.81
Excluding SY + GY + Family FE	0.00	0.05	-0.50	-0.08	-0.00	0.08	0.65
Excluding SY + GY + FY FE	0.00	0.04	-0.50	-0.07	-0.00	0.07	0.52

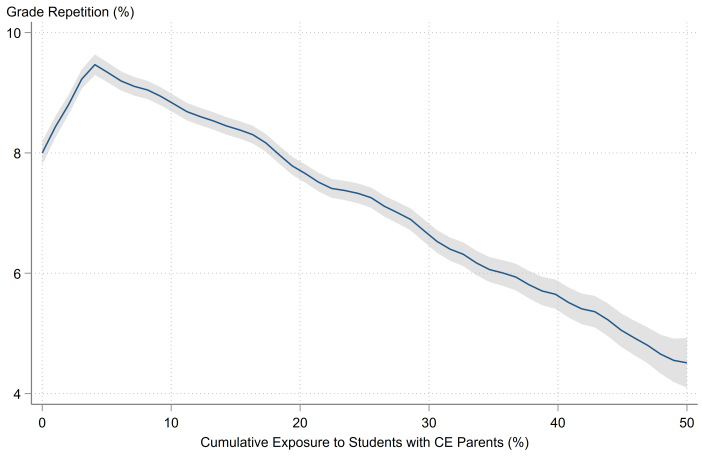
Notes. The table presents summary statistics of cumulative exposure, sequentially partialing out for different sets of fixed effects. SY stands for school-by-year. GY stands for grade-by-year. FY stands for family-by-year. The table uses data from the main analysis sample, which includes all public school students in mainland Portugal that can be followed since Grade 1, between 2006 and 2016, and have at least one sibling enrolled in the public education system in the same academic year. The summary statistics of this sample can be found in Table 1.

Figure A.2. Distribution of Family Size in the Full Sample



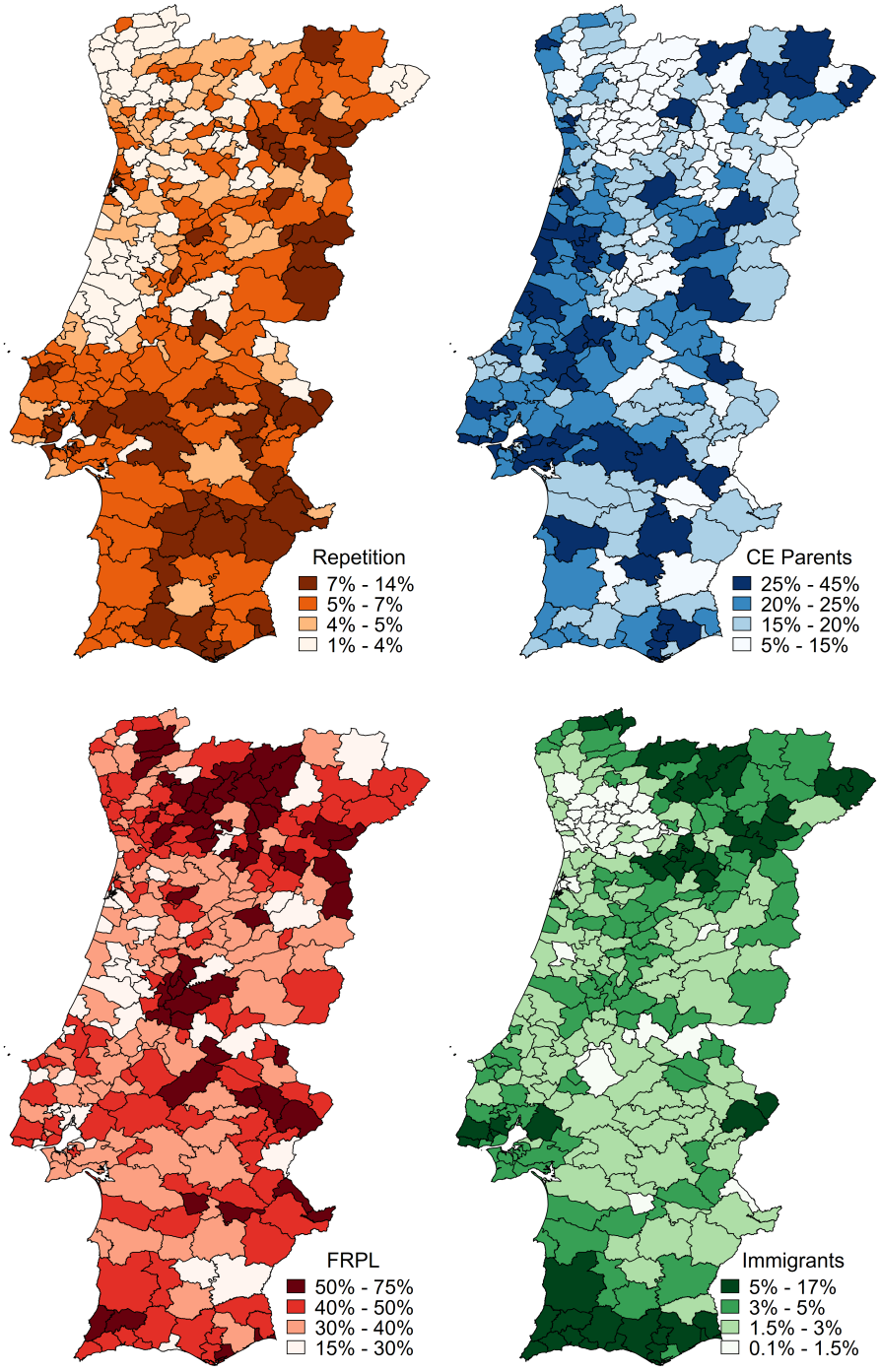
Notes. The figure depicts the distribution of the number of children in the household. The unit of observation is the family. The sample includes all public school students in mainland Portugal that can be followed since Grade 1, between 2006 and 2016.

Figure A.3. The Relationship between Grade Repetition and Cumulative Exposure to Students with CE Parents



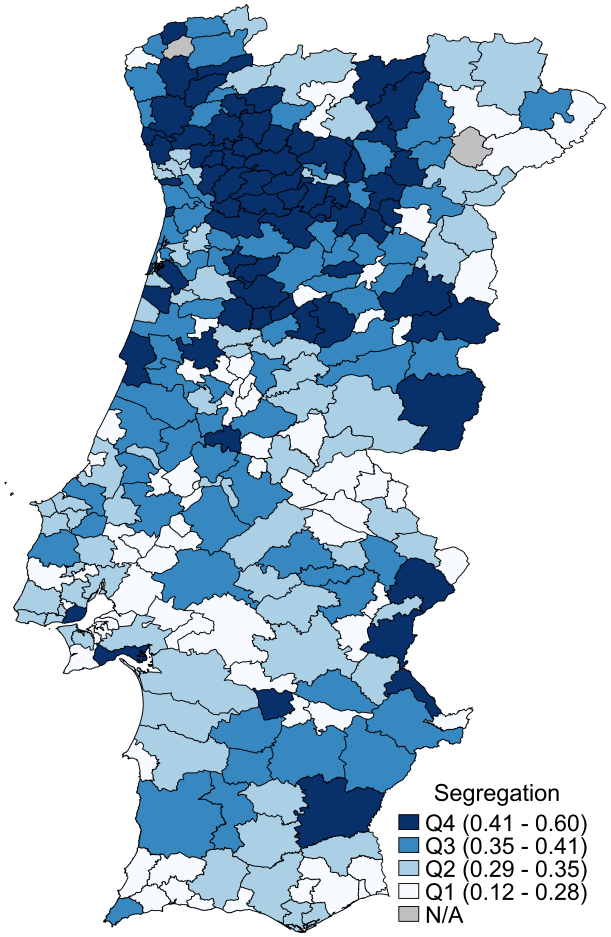
Notes. The figure depicts estimates of the relationship between grade repetition and cumulative exposure to students with college-educated parents using a local-linear estimator The sample to produce the estimates is the main analysis sample, including all students without college-educated parents, with at least one sibling in the same academic year, enrolled in a public school. The solid line represents the point estimates. The shaded area represents the 95% confidence interval. The estimates are computed for the range between 0 and 50% of cumulative exposure, which includes over 99% of the observations.

Figure A.4. The Geography of Student Characteristics



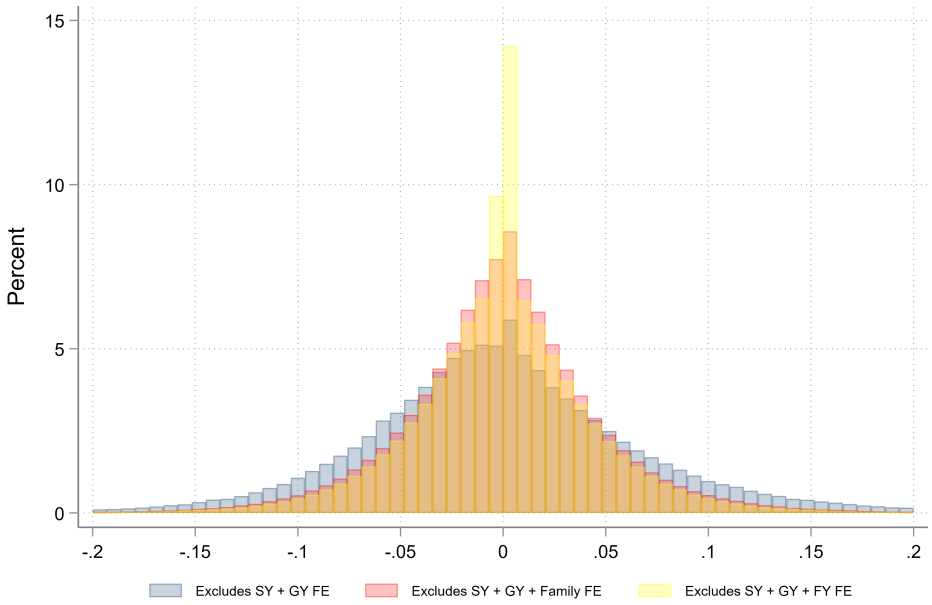
Notes. The figure depicts the administrative frontiers of mainland Portugal municipalities. The maps illustrate the percentage of grade repeaters, college-educated parents, those eligible for free or reduced price lunch (FRPL) and immigrants within each municipality. The sample includes all students enrolled in a public school.

Figure A.5. Within-Municipality, Across-School Segregation by Parental Education



Notes. The figure depicts the administrative frontiers of mainland Portugal municipalities. The map illustrates within-municipality, across school segregation by parental background. The segregation index is computed at the cohort level. For each municipality-grade-year (mgt) cell we compute a dissimilarity index given by $Seg_{mgt} = \frac{1}{2}\sum_{s\in mgt}\left|\frac{\text{CE}_s}{\text{CE}_{mgt}} - \frac{\text{NCE}_s}{\text{NCE}_{mgt}}\right|$. The index is then averaged at the municipality level, weighting for the number of schools in each municipality. As $Seg_{mgt} \to 1$, children with college-educated (CE) parents tend to be concentrated in just one school. As $Seg_{mgt} \to 0$, students with CE parents are evenly distributed across schools. Different colors represent different quartiles of the level of segregation, presented in the legend. The sample includes all students enrolled in a public school.

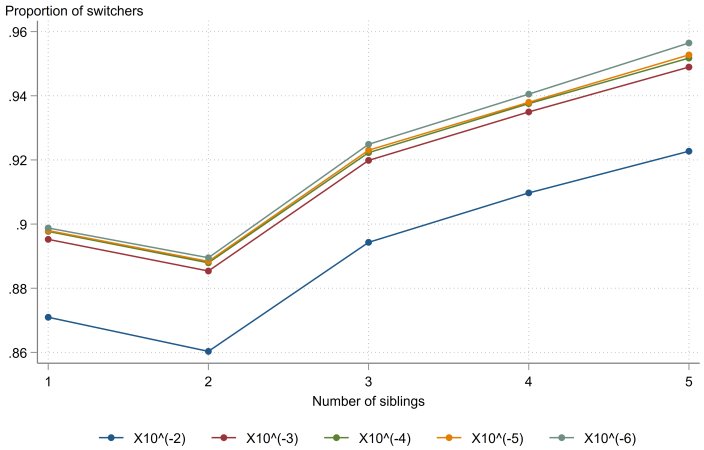
Figure A.6. Distributions of Residual Variation of Exposure for Models with Different Sets of Fixed Effects



Notes. The figure depicts the distribution of cumulative exposure to students with college-educated parents, as defined in Equation 1 of the main text, residualized from different sets of fixed effects. SY stands for school-by-year. GY stands for grade-by-year. FY stands for family-by-year.

B. Selection into Identification

Figure B.7. Proportion of Switchers by Definition of Switching and Number of Siblings in the Family



Notes. The figure depicts the proportion of switchers, defined as the individuals for which there is non-zero variation in the residualized measure of cumulative exposure to students with college-educated parents, for different definitions of null variation and by number of siblings in the family.

Table B.9. Summary Statistics by Switcher Status

	(1) Switchers	(2) Non-Switchers	(3) Difference (2) - (1)
Repeat the Grade (%)	8.0	5.7	-2.26
	(27.1)	(23.2)	(0.10)
Cumulative Exposure to Students with CE Parents	0.2	0.2	-0.01
	(0.1)	(0.1)	(0.00)
Contemporaneous Exposure to Students with CE Parents	0.2	0.2	-0.01
	(0.1)	(0.1)	(0.00)
Cumulative Exposure to FRPL Students	0.4	0.4	-0.03
	(0.2)	(0.2)	(0.00)
Cumulative Exposure to Immigrant Students	0.0	0.0	0.00
	(0.1)	(0.1)	(0.00)
Female (%)	48.3	51.2	2.87
	(50.0)	(50.0)	(0.19)
Age (in Years)	9.9	9.0	-0.88
	(2.6)	(2.3)	(0.01)
Free or Reduced Price Lunch [FRPL] (%)	59.8	55.4	-4.37
	(49.0)	(49.7)	(0.18)
Immigrants	1.9	2.2	0.32
	(13.8)	(14.8)	(0.05)
Internet at Home (%)	52.1	46.9	-5.21
	(50.0)	(49.9)	(0.19)
Number of Siblings	1.7	1.6	-0.10
	(0.9)	(0.8)	(0.00)
Birth Order	1.9	1.5	-0.40
	(0.9)	(0.7)	(0.00)
Observations	687,450	78,604	766,054

Notes. The table presents summary statistics of the main variables, by switcher status. Switchers are those for which there is non-zero variation in the residualized measure of cumulative exposure to students with college-educated parents, for the model with school-by-year, grade-by-year, and family-by-year fixed effects. Non-zero variation in this case is defined as all residuals r that are higher than $s.d.(r) \times 10^{-3}$ in absolute value, where $s.d.(r)$ is the standard deviation of the residuals. Non-switcher observations are defined as all other observations. Columns 1 and 2 present the means of each variable for switchers and never switchers, respectively, with standard deviations presented in parentheses, below. Column 3 presents the difference between the samples in Column 2 and 1, with the standard errors of a t-test presented in parantheses, below.

C. Conceptual Framework

C.1. Identification

We assume that each individual i has potential outcomes $y(e)$, where e is the potential level of exogenous cumulative exposure to peers with college-educated parents. We assume there is a linear potential outcome model with homogeneous causal parameters (η, τ), such that:

$$y_i(e) = \eta + \tau e. \tag{8}$$

In this framework, η identifies the outcome of individual i in the counterfactual where they would not be exposed to any peers with college-educated parents. The parameter τ identifies the causal, linear effect of an additional unit of exposure. The observed outcomes for each individual (Y_i) are defined at the observed level of cumulative exposure E_i, among other causes, such that $Y_i = Y_i(E_i, .)$. However, E_i is not exogenously assigned to each individual. Parents enroll their first child in first grade $(g = 1)$ of school s, in a given academic year t. The cohort is defined as the combination of school, year, and grade (sgt). Older children are subsequently enrolled in the following academic years, being exposed to different concentrations of peers with college-educated (CE) parents.

Table B.10. Cumulative Exposure to Students with CE Parents and Covariates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.0012	0.00081						
	(0.00033)	(0.00023)						
Free or Reduced Price Lunch			-0.039	-0.0049				
			(0.00048)	(0.00055)				
First Generation Immigrant					-0.0051	-0.0037		
					(0.0012)	(0.0012)		
Birth Order=2							0.020	0.0025
							(0.00055)	(0.00040)
Birth Order=3							0.0026	0.0053
							(0.00067)	(0.00074)
Birth Order=4							-0.0070	0.0065
							(0.00093)	(0.0011)
Birth Order=5							0.0051	0.012
							(0.0017)	(0.0017)
Birth Order=6							0.010	0.018
							(0.0038)	(0.0030)
R-squared	0.000	0.895	0.019	0.895	0.000	0.895	0.005	0.895
Observations	766,054	766,054	766,054	766,054	766,054	766,054	766,054	766,054
Mean Dep. Var.	.188	.188	.188	.188	.188	.188	.188	.188
Mean Dep. Var. (Ommited)	.187	.187	.211	.211	.188	.188	.179	.179
School-by-Year FE	No	Yes	No	Yes	No	Yes	No	Yes
Grade-by-Year FE	No	Yes	No	Yes	No	Yes	No	Yes
Family-by-Year FE	No	Yes	No	Yes	No	Yes	No	Yes

Notes. The table reports OLS estimates of regressions of cumulative exposure to peers with college-educated parents on a series of observable individual covariates. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The standardized coefficient shows the effect of a standard deviation change in the independent variable in terms of standard deviations of the dependent variable. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Our identifying assumption states that differences in exposure across siblings, in the same year, partialled out of variation across schools and trends for each grade across time, are unconfounded.

Assumption 1 *(Family-level Unconfoundedness). Given the causal model in Equation 8, and observed cumulative exposure to peers with college-educated parents, E_{isgt}, for each individual i, in school s, grade g, and year t,*

$$(\eta, \tau) \perp\!\!\!\perp E_{isgt} | \lambda_{ft(i)}, \gamma_{gt(i)}, \theta_{st(i)},$$

where λ_{ft} identifies the family f of individual i in year t, γ_{gt} identifies the grade in which the individual is enrolled in year t, and θ_{st} identifies the school in which the individual is enrolled in year t.

Assumption 1 implies that the identifying variation in the data comes mostly from either siblings

enrolled in different grades in the same school, or siblings enrolled in the same grade but in different schools, in the same academic year.

C.2. Within-Family Spillovers

To the extent that there are peer effects of exposure to higher parental education at the school level, it seems implausible there would not be effects of this exposure running between siblings. For instance, recent empirical literature has been demonstrating the influence of first-born children on the educational choices and outcomes of their siblings (e.g. Altmejd et al., 2021). Given our research design, we cannot separately identify the direct effect of peer exposure from the indirect, spillover effect of siblings exposure. In the presence of intrahousehold spillover effects, we require additional identifying assumptions.

To see this more formally, consider an extended causal model, relative to the one in Equation 8, where potential outcomes depend, linearly, on the joint distribution of own (e) and average siblings exposure ($\bar{e}_f$), with causal parameters (η, τ, ζ):

$$y_i(e, \bar{e}_f) = \eta + \tau e + \zeta \bar{e}_f. \tag{9}$$

What can within-family comparisons identify? As an illustration, consider the case of a family with only two siblings (i and j) and suppose that e and $\bar{e}_f$ are exogenously awarded. The difference in outcomes between the two siblings will be given by: $y_i(e_i, e_j) - y_j(e_j, e_i) = (\tau - \zeta) \times (e_i - e_j)$. By comparing the outcomes of the two siblings we cannot separately identify the direct effect (τ) from the intrahousehold spillover (ζ). We make the following two assumptions:

Assumption 2 *(Monotonic Interference). Given the causal model in Equation 9, suppose that sign*$(\tau) = s$*, then sign*$(\zeta) = s$*, for all* $s \in \{-1, 0, 1\}$.

Assumption 3 *(Sufficiently Weak Interference). Given the causal model in Equation 9 and Assumption 2, then* $|\zeta| \leq |\tau|$.

Taken together, Assumptions 2 and 3 mean that the exogenous, indirect effect of siblings exposure on an individual's outcome (ζ) has the same sign, but smaller magnitude than the direct effect (τ). From these two restrictions on intrahousehold spillovers we derive the following proposition:

Proposition 1 *(Bounded Effect). Given the causal model in Equation 9, suppose that sign*$(\tau) = s, \forall s \in \{-1, 0, 1\}$*, and that Assumptions 2 and 3 hold, then* $|\tau| \geq |\tau - \zeta| \geq 0$.

Proposition 1 implies that, in the presence of intrahousehold spillovers, we can at most identify a lower bound of the direct effect by making within-family comparisons, as long as the indirect effect is sufficiently small and of the same direction as the direct effect. Therefore, the validity of our interpretation of the estimated β, from Equation 2, as a lower bound of τ rests upon the plausibility of Assumptions 2 and 3.

To test the relative importance of intrahousehold spillovers we run the following specifications, on all individuals, and separately for individuals of each birth order:

$$Y_{isgt} = \mu + \gamma_{gt} + \theta_{st} + \rho E_{isgt} + \phi \bar{E}^{-i}_{isgt} + \mathbb{X}_{isgt} \cdot \delta + \epsilon_{isgt}, \tag{10}$$

where $\bar{E}^{-i}_{isgt}$ is the average cumulative exposure of individual i siblings to peers with college-educated parents, in year t. Notice that, in contrast to the main specification defined in Equation 2, we do not include family-by-year fixed effects, as it would not allow us to estimate ρ and ϕ separately. Without interpreting either of these estimates as causal, we use the specification in Equation 10, to claim that $sign(\hat{\rho}) = sign(\hat{\phi})$ and $|\hat{\rho}| \geq |\hat{\phi}|$ lends credibility to the idea that we can identify a lower bound of our peer effect of interest.

Table C.1. Own and Siblings Cumulative Exposure to Peers with CE Parents

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Full Sample		First Borns		Second Borns		Third Borns	
Exposure to CE	-0.106	-0.088	-0.156	-0.126	-0.082	-0.068	-0.123	-0.105
	(0.0044)	(0.0043)	(0.0085)	(0.0084)	(0.0063)	(0.0063)	(0.0152)	(0.0150)
Siblings Exposure to CE	-0.065	-0.045	-0.063	-0.050	-0.041	-0.033	-0.029	-0.017
	(0.0034)	(0.0034)	(0.0057)	(0.0057)	(0.0059)	(0.0058)	(0.0130)	(0.0129)
R-squared	0.109	0.121	0.126	0.138	0.161	0.171	0.273	0.281
Observations	765,845	765,845	288,068	288,068	325,387	325,387	90,544	90,544
Mean Dep. Var.	.077	.077	.078	.078	.071	.071	.089	.089
School-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	No	Yes	No	Yes	No	Yes	No	Yes

Notes. The table reports estimates of ρ and ϕ, for alternative specifications of the preferred model presented in Equation 10. The sample in Columns 1 and 2 is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The other samples restrict, respectively to first-born (Columns 3 and 4), second-born (Column 5 and 6) and third-born (Column 7 and 8) siblings. The dependent variable is grade repetition in the end of a given grade-year. The regressors of interest are the cumulative exposure to students with CE parents since Grade 1 until the current grade, as defined in Equation 2, as well as an average of the same measure among the siblings in the household, excluding the own individual. The standardized coefficient shows the effect of a standard deviation change in the independent variable in terms of standard deviations of the dependent variable. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Table C.1 reports estimates of ρ and ϕ, in Equation 10 for four different samples. The first sample includes all siblings enrolled in public school in the same academic year and whose parents are not college-educated (Columns 1 and 2). The other samples restrict, respectively, to first-born (Columns 3 and 4), second-born (Column 5 and 6) and third-born (Column 7 and 8) siblings. All specifications include grade-by-year and school-by-year fixed effects. As hypothesized, across all samples, we verify that: (i)

own exposure, as well as siblings exposure have the same direction of effect on grade repetition; and that (ii) the estimated exposure of siblings has a lower effect than own exposure to peers with CE parents. The qualitative interpretation of these coefficients is robust to controlling for individual covariates.

D. Classroom Exposure

D.1. Endogeneity of Classroom Formation

In this section we show that students are not randomly allocated to classes within schools. We proceed in two ways. First, we implement the test in Ammermueller and Pischke (2009). For the subsample of schools with two or more classes, we run Pearson χ^2 tests to investigate whether there are more students with CE parents in given classes than what would be consistent with independence, given the number of such students in the cohort. We reject the null of random allocation ($p < 0.01$).

Second, consistent with this test, we also report substantial within-cohort, across-classroom segregation on the basis of parental education, using the dissimilarity index proposed by Allen et al. (2015), which adapts the Duncan and Duncan (1995) segregation index:

$$\text{Seg}_{sgt} = \frac{1}{2} \sum_{c \in sgt} \left| \frac{\text{CE}_c}{\text{CE}_{sgt}} - \frac{\text{NCE}_c}{\text{NCE}_{sgt}} \right|, \tag{11}$$

where CE_c is the number of students with CE parents in class c, CE_{sgt} is the number of students with CE parents in cohort sgt, and NCE_c and NCE_{sgt} are analogous measures for the case of students without college educated parents. Seg_{sgt} measures how fairly distributed are students of both groups distributed across classes within a given cohort. As $\text{Seg}_{sgt} \to 0$, the proportion of students with college-educated parents is identical across classes. As $\text{Seg}_{sgt} \to 1$, the higher the concentration of students with college-educated parents in the same classroom.

We use the bias-corrected version of the dissimilariy index in Equation 11, as in Allen et al. (2015). The need for correction stems from the fact that for certain combinations of cohort and group sizes, the dissimilarity index above will overestimate the level of segregation. For instance, if there are too few students of a given group in a cohort, their concentration in only one class is much more likely to be due to chance than to discretionary decisions at the school level. Therefore, for schools where segregation cannot be distinguished from random chance, the bias-corrected version of the measure in Equation 11 is conservative and sets the level of segregation to zero. Still, we observe a mean across-classroom segregation of 0.2, with a standard deviation of 0.19.

D.2. Classroom Exposure

In this section, we reproduce the analysis of section 4 using the classroom exposure instead of cohort-level exposure as defined in Equation 1. For each student i, in school s, class c, and year t, classroom level exposure is defined as:

$$E_{isct} = \frac{1}{t - \underline{t} + 1} \sum_{\underline{t} \leq t' \leq t} \frac{\#\text{ Peers with CE Parents}_{isct'}}{\#\text{ Peers}_{isct'}}. \tag{12}$$

Table D.1 reports results for a series of different specifications. Columns 1 through 4 reproduce the

specifications in Table 2, using exposure at class level. The coefficient of −0.11 implies that a standard deviation increase in cumulative classroom exposure (13.8 p.p.) is associated with a 1.4 p.p. decrease in grade repetition, or 18.5 % over the sample mean. The coefficient associated with classroom exposure is 1.4 times larger in magnitude than the one estimated for cumulative cohort exposure (Table 2, Column 4). The two coefficients are also statistically different from each other as the 95% confidence intervals of each do not include the point estimate of the other.

Table D.1. Grade Repetition and Classroom Exposure to Students with College-Educated Parents

	Outcome: Grade Repetition						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Exposure to CE	-0.109	-0.140	-0.041	-0.041	-0.073	-0.075	-0.057
	(0.0023)	(0.0032)	(0.0047)	(0.0047)	(0.0130)	(0.0130)	(0.0074)
R-squared	0.004	0.110	0.610	0.611	0.810	0.811	0.611
Observations	766,054	766,054	766,054	766,054	621,472	621,472	766,054
Mean Dep. Var.	.077	.077	.077	.077	.077	.077	.077
School-by-Year FE	No	Yes	Yes	Yes	No	Yes	Yes
Grade-by-Year FE	No	Yes	Yes	Yes	No	Yes	Yes
Family-by-Year FE	No	No	Yes	Yes	Yes	Yes	Yes
Class FE	No	No	No	No	Yes	Yes	No
Individual Controls	No	No	No	Yes	No	Yes	Yes

Notes. The table reports estimates of β, for alternative specifications of the preferred model presented in Equation 2. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the cumulative classroom exposure to students with CE parents since Grade 1 until the current grade, as defined in Equation 2. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

In Columns 5 and 6, we control for differences across classes within school, by including class fixed effects. In Column 5, we include family-by-year fixed effects, while in Column 6 we further control for grade-by-year and school-by-year fixed effects. We estimate a coefficient of −0.075 or 2.6 times larger than the point estimates of our preferred specification in Table 2. Finally, Column 7 includes the cohort exposure to CE parents as a control. We find that the classroom exposure matters for grade repetition beyond the cohort level of exposure.

E. Additional Results

E.1. Descriptives

Table E.1. Association between Eligibility to Repeat the Grade and Grade Repetition

	Outcome: Grade Repetition			
	(1)	(2)	(3)	(4)
Eligible to Repeat	0.588	0.587	0.591	0.603
	(0.0024)	(0.0024)	(0.0024)	(0.0036)
Eligible to Repeat × CE Parents			-0.049	-0.041
			(0.0052)	(0.0053)
R-squared	0.549	0.549	0.549	0.550
Observations	1,741,750	1,741,750	1,741,750	1,741,750
Mean Dep. Var.	.071	.071	.071	.071
Mean Dep. Var. (Ineligible)	.001	.001	.001	.001
Grade-by-Year FE	Yes	Yes	Yes	Yes
Individual Controls	No	Yes	No	Yes

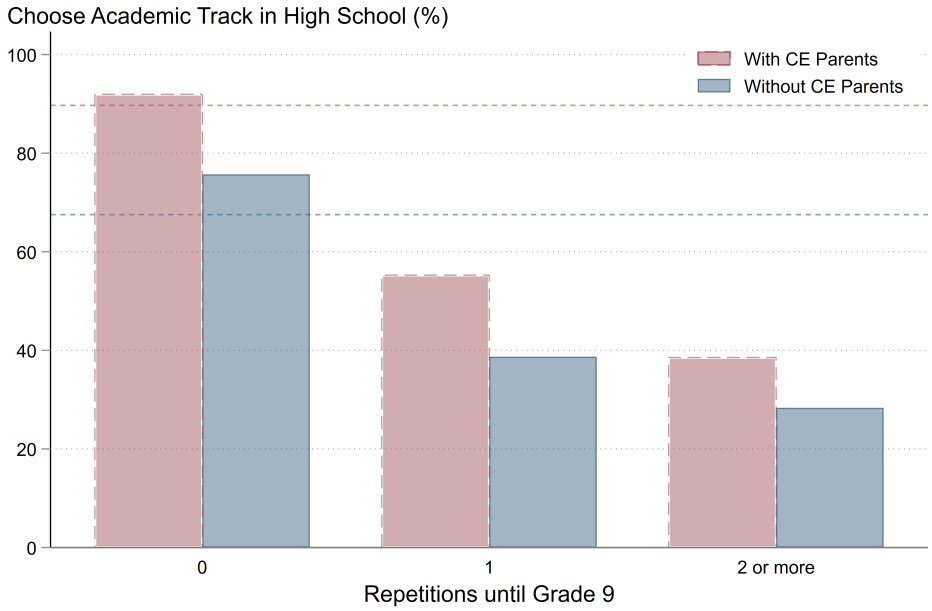
Notes. The table presents coefficients of OLS regressions. The sample includes all students that can be followed since Grade 1, enrolled between Grades 5 and 9, for which there is information on their school grades. The dependent variable on all specifications is a dummy variable indicating whether the individual repeated the grade in a given year. The independent variable of interest in Columns 1 and 2 is a dummy indicating whether the student is eligible to repeat the grade according to the criteria described in the main text. Columns 3 and 4 also include a dummy identifying whether the individual has college-educated (CE) parents, as well as an interaction term between this variable and eligibility. Individual controls include indicator variables identifying whether the student has college-educated parents, is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Table E.2. Association between Having College-Educated Parents and Main Variables

	(1)	(2)	(3)	(4)	(5)	(6)
Outcomes	Repeat	Cumulative Exposure	GPA	Fail Language	Fail Math	Prop. Failed Subjects
CE Parents	-0.050	0.124	0.551	-0.108	-0.222	-0.087
	(0.0003)	(0.0007)	(0.0017)	(0.0007)	(0.0009)	(0.0004)
R-squared	0.030	0.157	0.137	0.027	0.061	0.057
Observations	5,053,059	5,053,059	1,779,230	1,760,107	1,767,521	1,779,230
Mean Dep. Var.	.051	.219	3.507	.124	.27	.099
Mean Dep. Var. (No CE Parents)	.063	.19	3.387	.148	.318	.118
Grade-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes

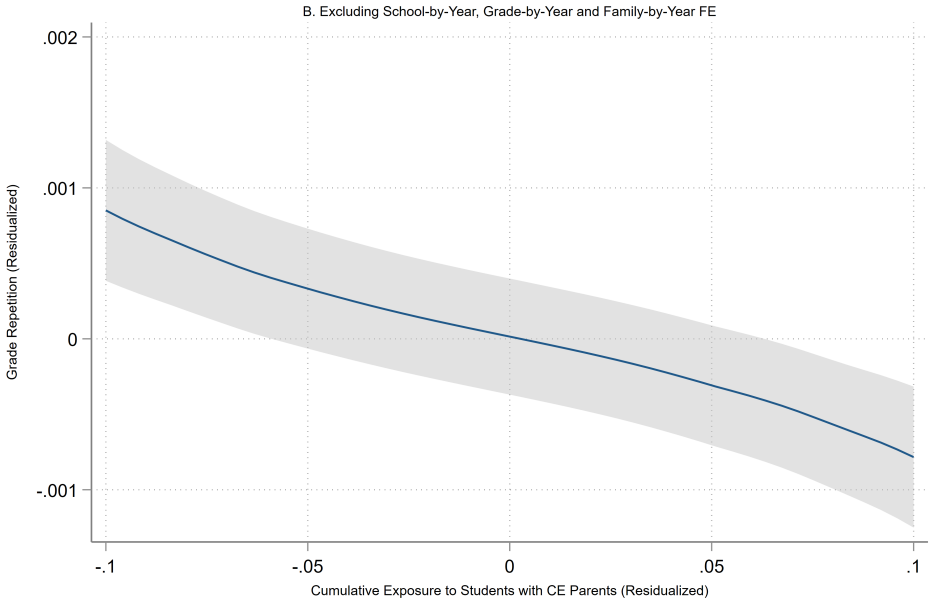
Notes. The table presents coefficients of OLS regressions. The sample includes all students that can be followed since Grade 1, enrolled between Grades 1 and 9. The dependent variables are presented under each column identifies. The independent variable odummy indicating whether the student has at least one college-educated parent. All specifications control for grade-by-year-fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Figure E.1. Choice of Academic Track in High School, by Parental Education and Repetitions until Grade 9



Notes. The figure depicts the percentage of students who choose the academic track in high school, as opposed to the vocational track, by parental education and the number of grade repetitions until Grade 9. The dash lines represent the percentage of students that choose the academic track in high school, respectively for those with and without at least one college-educated parent. The sample includes all students that can be followed from Grade 1 until Grade 10.

Figure E.2. Relationship between Exposure and Grade Repetition under a Model with Family-by-Year Fixed Effects



Notes. The figure depicts estimates of the relationship between grade repetition and cumulative exposure to students with college-educated parents, partialled out of school-by-year, grade-by-year, and family-by-year fixed effects, using a local-linear estimator. Residuals are computed using the main analysis sample, including all students without college-educated parents, with at least one sibling in the same academic year, enrolled in a public school. The solid line represents the point estimates. The shaded area represents the 95% confidence interval. The estimates are computed for the range between −.1 and .1 of residualized cumulative exposure, which includes over 90% of observations.

E.2. Robustness

Table E.3. Robustness to Alternative Measure of Exposure

	Outcome: Grade Repetition			
	(1)	(2)	(3)	(4)
Exposure to CE (Alternative Measure)	-0.097	-0.130	-0.032	-0.031
	(0.0027)	(0.0046)	(0.0065)	(0.0065)
R-squared	0.002	0.109	0.610	0.611
Observations	766,054	766,054	766,054	766,054
Mean Dep. Var.	.077	.077	.077	.077
School-by-Year FE	No	Yes	Yes	Yes
Grade-by-Year FE	No	Yes	Yes	Yes
Family-by-Year FE	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

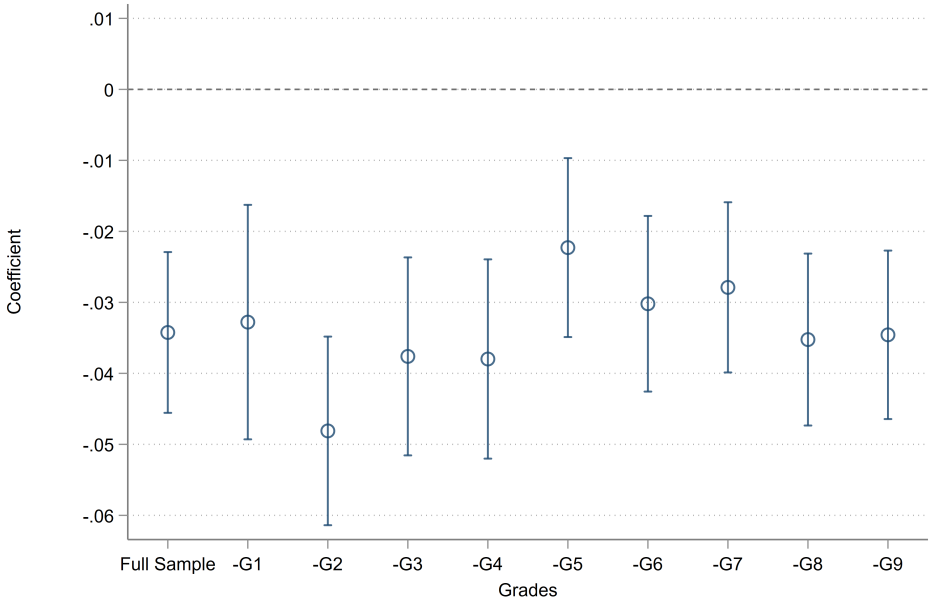
Notes. The table shows regression coefficients analogous to those of Table 2 in the main text. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the cumulative exposure to students with CE parents since Grade 1 until the current grade, as defined in Equation 2, in the main text. For this alternative measure, students with missing information on parental education are treated as not having college-educated parents. The standardized coefficient shows the effect of a standard deviation change in the independent variable in terms of standard deviations of the dependent variable. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Table E.4. Results with Contemporaneous Measure of Exposure

	Outcome: Grade Repetition			
	(1)	(2)	(3)	(4)
Exposure to CE (Contemporaneous Measure)	-0.064	-0.035	-0.019	-0.018
	(0.0027)	(0.0045)	(0.0066)	(0.0066)
R-squared	0.001	0.108	0.610	0.611
Observations	765,698	765,597	765,309	765,309
Mean Dep. Var.	.077	.077	.077	.077
School-by-Year FE	No	Yes	Yes	Yes
Grade-by-Year FE	No	Yes	Yes	Yes
Family-by-Year FE	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes. The table shows regression coefficients analogous to those of Table 2 in the main text. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the contemporaneous exposure to students with CE parents, defined as the percent of students with CE parents that a student is exposed to in a given school-grade-year cell. For this alternative measure, students with missing information on parental education are treated as not having college-educated parents. The standardized coefficient shows the effect of a standard deviation change in the independent variable in terms of standard deviations of the dependent variable. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Figure E.3. Estimated Effects in the Analysis Sample and in Samples Excluding Each Grade



Notes. The figure depicts estimates of β in Equation 2 in the main text, for different samples. Bars represent 95% confidence intervals. Each presented sample -Gj, with $j = 1, ..., 9$ represents the full sample for which main results are presented in Table 2, excluding children when they are observed at grade j.

Table E.5. Results with Absolute Exposure

	Outcome: Grade Repetition			
	(1)	(2)	(3)	(4)
Exposure to CE (Absolute)	-0.00013	-0.00132	-0.00054	-0.00051
	(0.000)	(0.0001)	(0.0001)	(0.0001)
R-squared	0.000	0.108	0.610	0.611
Observations	766,054	766,054	766,054	766,054
Mean Dep. Var.	.077	.077	.077	.077
School-by-Year FE	No	Yes	Yes	Yes
Grade-by-Year FE	No	Yes	Yes	Yes
Family-by-Year FE	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes. The table shows regression coefficients analogous to those of Table 2 in the main text. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is cumulative exposure, defined as the average number of peers with CE parents that a student is exposed to. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Table E.6. Results with Alternative Specification

	Outcome: Grade Repetition							
	Full Sample		First Borns		Second Borns		Third Borns	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exposure Difference to Siblings	-0.049	-0.046	-0.019	-0.020	-0.061	-0.061	-0.109	-0.109
	(0.0037)	(0.0037)	(0.0065)	(0.0064)	(0.0059)	(0.0059)	(0.0140)	(0.0139)
R-squared	0.045	0.047	0.083	0.085	0.112	0.113	0.225	0.226
Observations	765,845	765,845	288,068	288,068	325,387	325,387	90,544	90,544
Mean Dep. Var.	0	0	.017	.017	-.009	-.009	-.014	-.014
School-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	No	Yes	No	Yes	No	Yes	No	Yes

Notes. The table shows regression coefficients using an alternative specification to the one in main specification in Equation 2. Instead of relying on family-by-year fixed effects, we subtract from each individual's exposure and outcome variable, the average of their siblings for the same variables. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is de-meaned grade repetition in the end of a given grade-year. The regressor of interest is de-meaned cumulative exposure to students with CE parents, defined as the percent of students with CE parents that a student is exposed to in a given school-grade-year cell. In Columns 3 through 8, the same specifications are presented separately by those who are first born, second born and third born. Individual controls include indicator variables identifying whether the individual is female, foreign-born, has internet at home, and benefits from FRPL, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Table E.7. Alternative Level of Clustered Standard Errors

Cluster Level	Standard Error	C.I. Lower Bound	C.I. Upper Bound
School-by-Grade-by-Year	0.0063	-0.0410	-0.0165
Individual	0.0059	-0.0403	-0.0172
Family	0.0058	-0.0400	-0.0175

Notes. The table shows robust clustered standard errors and respective confidence intervals associated with estimates of β, according to the specification in Equation 2 in the main text.

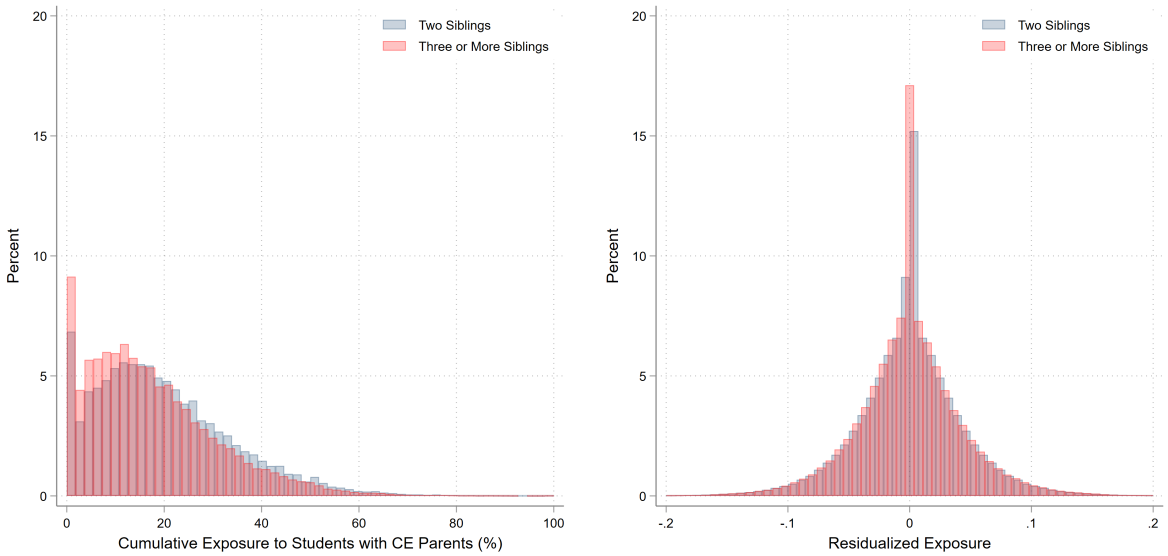
E.3. Heterogeneity

Table E.8. Heterogeneity by Family Size

	Two Siblings			Three or More Siblings		
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure to CE	-0.096	-0.010	-0.009	-0.149	-0.051	-0.052
	(0.0053)	(0.0078)	(0.0078)	(0.0076)	(0.0112)	(0.0112)
Std. Coefficient	-.056	-.006	-.005	-.067	-.023	-.024
R-squared	0.124	0.642	0.644	0.165	0.621	0.623
Observations	423,232	421,471	421,471	338,419	336,016	336,016
Mean Dep. Var.	.062	.062	.062	.096	.096	.096
School-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Grade-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Family-by-Year FE	No	Yes	Yes	No	Yes	Yes
Individual Controls	No	No	Yes	No	No	Yes

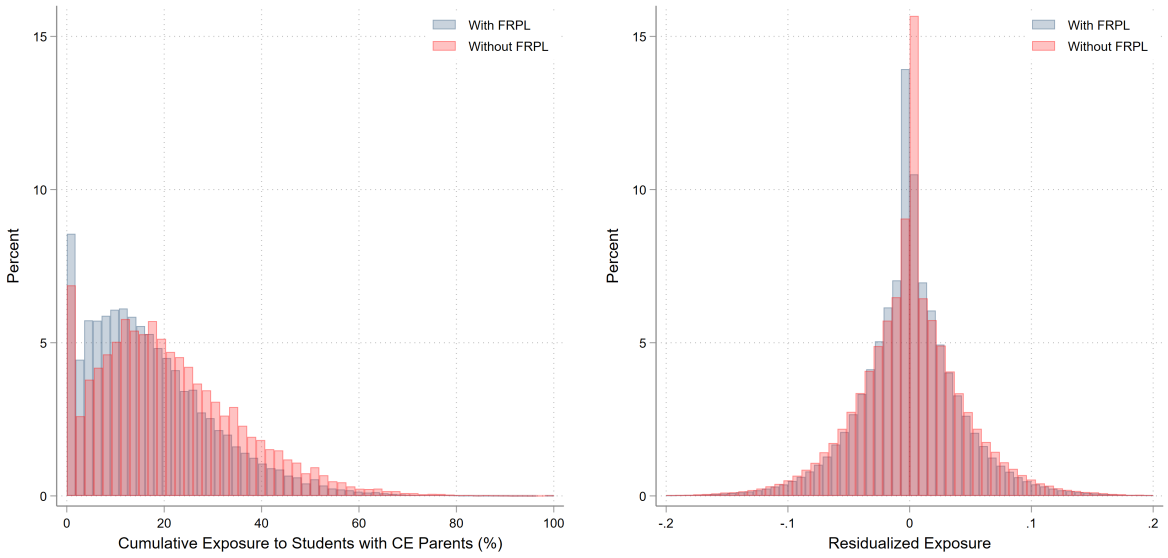
Notes. The table shows regression coefficients analogous to those of Table 2 in the main text. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the cumulative exposure to students with CE parents since they are first enrolled in Grade 1. In Columns 1 through 3, the sample only includes families with two children. In Columns 4 through 6, the sample only includes families with more than two children. The standardized coefficient shows the effect of a standard deviation change in the independent variable in terms of standard deviations of the dependent variable. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Figure E.4. Distributions of Cumulative Exposure by Family Size



Notes. The figure depicts histograms of cumulative exposure by family size. The left-hand side panel presents the distribution for the raw variation in cumulative exposure to students with college-educated parents. The right-hand side panel presents the distribution for the variation in cumulative exposure residualized of school-by-year, grade-by-year and family-by-year fixed effects.

Figure E.5. Distributions of Cumulative Exposure by FRPL



Notes. The figure depicts histograms of cumulative exposure by free or reduced price lunch (FRPL) status. The left-hand side panel presents the distribution for the raw variation in cumulative exposure to students with college-educated parents. The right-hand side panel presents the distribution for the variation in cumulative exposure residualized of school-by-year, grade-by-year and family-by-year fixed effects.

Table E.9. Heterogeneity by Siblings Gender Mix

	Outcome: Grade Repetition					
	Males		Females		Mixed Gender	
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure to CE	-0.144	-0.0387	-0.0994	-0.00724	-0.136	-0.0286
	(0.00985)	(0.0151)	(0.00893)	(0.0139)	(0.00629)	(0.00924)
R-squared	0.197	0.674	0.213	0.694	0.141	0.614
Observations	193,387	193,387	171,923	171,923	380,408	380,408
Mean Dep. Var.	.088	.088	.06	.06	.08	.08
School-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Grade-by-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Family-by-Year FE	No	Yes	No	Yes	No	Yes
Individual Controls	No	Yes	No	Yes	No	Yes

Notes. The table shows regression coefficients analogous to those of Table 2 in the main text. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year and whose parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the cumulative exposure to students with CE parents since they are first enrolled in Grade 1. In Columns 1 through 2, the sample only includes families with only boys. In Columns 3 and 4 the sample only includes families with only girls. In Columns 5 and 6 the sample only includes families with a mix of boys and girls. Individual controls include indicator variables identifying whether the is female, foreign-born, has internet at home, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Table E.10. Results Including Families with College-Educated Parents

	Outcome: Grade Repetition			
	(1)	(2)	(3)	(4)
Exposure to CE	-0.137	-0.148	-0.029	-0.027
	(0.0016)	(0.0030)	(0.0039)	(0.0039)
R-squared	0.009	0.095	0.604	0.605
Observations	1,174,502	1,174,502	1,174,502	1,174,502
Mean Dep. Var.	.061	.061	.061	.061
School-by-Year FE	No	Yes	Yes	Yes
Grade-by-Year FE	No	Yes	Yes	Yes
Family-by-Year FE	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes. The table shows regression coefficients analogous to those of Table 2 in the main text. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year, including both those with and without parents have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the cumulative exposure to students with CE parents since they are first enrolled in Grade 1. Individual controls include indicator variables identifying whether the individual is a female, foreign-born, has internet at home, parental education, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.

Table E.11. Results Only Among Families with College-Educated Parents

	Outcome: Grade Repetition			
	(1)	(2)	(3)	(4)
Exposure to CE	-0.011	-0.014	-0.010	-0.010
	(0.0009)	(0.0023)	(0.0034)	(0.0034)
Std. Coefficient	-.022	-.029	-.02	-.02
R-squared	0.000	0.114	0.611	0.612
Observations	282,276	282,276	282,276	282,276
Mean Dep. Var.	.008	.008	.008	.008
School-by-Year FE	No	Yes	Yes	Yes
Grade-by-Year FE	No	Yes	Yes	Yes
Family-by-Year FE	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes. The table shows regression coefficients analogous to those of Table 2 in the main text. The sample is an unbalanced panel of students enrolled in public primary and middle schools in mainland Portugal, between Grades 1 and 9, who have at least a sibling enrolled in public school in the same academic year, with parents who have no college-education (CE). The dependent variable is grade repetition in the end of a given grade-year. The regressor of interest is the cumulative exposure to students with CE parents since they are first enrolled in Grade 1. Individual controls include indicator variables identifying whether the individual is a female, foreign-born, has internet at home, as well as birth order fixed effects. Robust standard errors clustered at the school-by-grade-by-year level are presented in parentheses.