UNIVERSITY OF MANNHEIM

University of Mannheim / Department of Economics

Working Paper Series

Gender Peer Effects in School, a Birth Cohort Approach

Antonio Ciccone Walter Garcia-Fontes

Working Paper 14-19

July 2014

Gender Peer Effects in School, a Birth Cohort Approach

Antonio Ciccone Walter Garcia-Fontes[1]

July 2014

[1]University of Mannheim and Barcelona GSE, and Universitat Pompeu Fabra and Barcelona GSE, respectively. We gratefully acknowledge the comments of Brindusa Anghel, Ghazala Azmat, Antonio Cabrales, Caterina Calsamiglia, Gabrielle Fack, Libertad Gonzalez, Stephan Litschig, Luis Ricardo Maertens, Marco Manacorda, Barbara Petrongolo, Joaquim Silvestre, Tetyana Surovtseva, Alessandro Tarozzi, Michele Tertilt, and seminar participants in Barcelona, Madrid, Mannheim, Munich, and London. The main data used have become available thanks to the efforts of Antonio Cabrales, Ismael Sanz Labrador, and Pablo Vazquez. We thank Luis Pires Jimenez for providing important additional data. We gratefully acknowledge financial support from CREI, fedea, Fundacio Catalunya - La Pedrera, Spanish research grants ECO2011-25272, ECO2011-30323-C03-02, and SEV-2011-0075 (Severo Ochoa Program for Centers of Excellence in R&D).

Abstract

We propose estimating gender peer effects in school by exploiting within-school variation in gender composition across birth cohorts. Our approach differs from the existing literature, which exploits variation in gender composition at a given grade level in different years. We argue that the birth cohort approach is a useful alternative as the grade level approach generally yields spurious gender peer effects when there is grade retention. The birth cohort approach applied to primary schools in Spain indicates statistically significant positive gender peer effects of girls on boys' academic achievement and statistically insignificant effects of girls on girls' achievement.

1 Introduction

Do girls learn more together with girls instead of boys? And what about boys? Possible gender peer effects in learning have been debated since the introduction of mixed-gender education and would have to be taken into account in the design of school systems as well as in the policy response to the recent revival of single-sex schools (e.g. Hoxby, 2000; Whitmore, 2005; Lavy and Schlosser, 2011).[1] Early empirical work looked for gender peer effects across schools but could not deal with the selection of students with different skills into different schools. The best evidence that gender composition affects learning in school comes from Hoxby's (2000) and Lavy and Schlosser's (2011) studies for Texas and Israel respectively. These studies bypass the selection of students with different skills into different schools by examining the response of academic achievement to within-school differences in gender composition at a given grade level in different years.[2] Examining the consequences of such differences in gender composition is appealing as they partly reflect natural variation in the births of girls and boys in school catchment areas (Lavy and Schlosser, 2011).

The within-school grade level approach to gender peer effects may not always be immune to selection issues however. Most school systems allow for grade retention of academically weak students and children in the same birth cohort may therefore end up in different grades depending on their academic skills. For example, only three of twenty-seven European Union countries rule out grade retention in primary school (European Commission, 2011). According to the program for international student assessment (PISA, 2009a), the share of students retained at least once in primary school averages to 8 percent across OECD countries and to 7 percent across EU countries.[3] Retention rates in lower-secondary school are similar. Countries with comparatively high retention rates in primary school are Belgium and France (17 percent), Spain (11 percent), and the USA (10.6 percent).[4] Retention rates

[1] For an analysis of optimal (student) assignments in the presence of social spillovers see Graham (2011) and Graham, Imbens, and Ridder (2010, 2014).

[2] Whitmore (2005) studies the effect of gender composition at the classroom level on boys' and girls' academic achievement in kindergarten and school. She finds mostly statistically insignificant effects, maybe because of a relatively small sample size compared to Hoxby or Lavy and Schlosser.

[3] Retention rates in non-OECD countries appear to be higher, see Manacorda (2012) and UNESCO (2002). In comparing retention rates across countries, it should be taken into account that the length of primary schooling differs. Children may also be retained in kindergarten or their entry into school may be delayed if parents and teachers consider the child not ready for school.

[4] There is no PISA data on the Texas school system studied by Hoxby (2000). The Texas education agency (1999, 2011) reports the share of students retained by grade and year between 1994 and 2009. If no student was retained more than once, the data would imply that 13-16 percent of students were retained once in primary school. As some students are retained more than once in primary school, 13-16 percent is an upper bound to the share of students retained at least once in primary school. The discrepancy should be small however as according to PISA (2009a) only around 0.5 percent of US students were retained more than once in primary school (there are no such data for Texas).

in primary school are comparatively low in Canada and Israel (4.2 percent) and the UK (1.7 percent).

To understand the effects of grade retention on the grade level approach to gender peer effects we develop a theoretical model of a school system with grade retention that we can solve analytically. In our model, students with academic skills below a threshold are retained in a grade at some point during primary school. As a result, academically weak students end up in a lower grade than their academically stronger peers in the same birth cohort. Another important feature of the model is that the gender composition of birth cohorts and the skills of girls and boys in a birth cohort may be subject to exogenous shocks. The question we ask is if and why spurious gender peer effects in academic achievement may emerge when gender peer effects within schools are estimated at the grade level.

A main result of our theoretical model is that the grade level approach generally yields spurious gender peer effects in academic achievement even if grade level differences in gender composition were solely driven by exogenous shocks to gender composition at the birth cohort level. Exogenous shocks to the skills of girls and boys at the birth cohort level also translate into spurious gender peer effects at the grade level. The direction of the spurious gender peer effects depends on the impact of grade repetition on students' academic skills. If grade repetition improves retained students' academic skills, exogenous shocks to the skills of girls and boys at the birth cohort level lead to a spurious positive gender peer effect of girls on girls' academic achievement and a spurious negative gender peer effect of girls on boys' achievement. If students who have been retained in the past perform on average worse academically than non-retained students in the same grade, exogenous shocks to the gender composition of birth cohorts also lead to a spurious positive gender peer effect of girls on girls' academic achievement and a spurious negative gender peer effect of girls on boys' achievement.

Because of the limitations of the grade level approach to gender peer effects in school systems with grade retention, we propose estimating gender peer effects in academic achievement by exploiting within-school differences in gender composition across birth cohorts. Students are assigned to the same birth cohort if they should have started school in the same year according to the school system's enrollment rule. The birth cohort approach examines whether girls or boys in a birth cohort with a greater share of girls do better academically than students of the same gender in other birth cohorts in the same school. Gender peer effects estimated using the birth cohort approach are intention-to-treat effects. Their magnitude depends on the strength of gender peer effects in the classroom but also, for example, on the rate of grade retention as this affects the link between the share of girls in the average

student's birth cohort and the share of girls in the average student's classroom throughout primary school. A key feature of the birth cohort approach to gender peer effects is that it does not yield spurious gender peer effects when there is grade retention.[5]

We use the birth cohort approach to estimate gender peer effects in a sample of Spanish primary schools. Spain allows for grade retention in primary school and the rate of grade retention is above the OECD and EU average. An advantage of our primary school data compared to Hoxby's (2000) data for Texas and Lavy and Schlosser's (2011) data for Israel is that they allow estimating gender peer effects at the birth cohort level (as well as the grade level).[6] Also, mobility across primary schools appears to be low compared to Texas and even to Israel, and compliance with the primary school enrollment rule is high.

When we estimate gender peer effects using the birth cohort approach, we find a statistically significant, positive gender peer effect of girls on the academic achievement of boys. A 10-percentage-point increase in the share of girls in a birth cohort improves boys' overall academic achievement and their achievement in mathematics by around 3 to 4 percent of a standard deviation. (Girls do worse than boys on average and this effect can therefore not be explained by spillovers from high-skill to low-skill students.[7]) On the other hand, the effect of the share of girls in a birth cohort on the achievement of girls is statistically insignificant. The grade level approach tends to yield a different pattern of gender peer effects than the birth cohort approach. The difference between the birth cohort and the grade level approach is consistent with our theoretical model if (i) grade repetition improves retained students' skills but students retained in the past still do worse on average than non-retained students in the same grade and (ii) grade level differences in gender composition are mostly driven by shocks to gender composition or skills at the birth cohort level.

As rates of grade retention in Spain are above the OECD and EU average, it is natural to wonder whether the birth cohort and the grade level approach can yield different patterns of gender peer effects when retention rates are around or below the OECD or EU average. We examine this question in a series of counterfactual experiments based on our theoretical model. After calibrating the model to match our gender-peer-effect estimates for Spain,

[5]It is natural to wonder whether spurious gender peer effects at the grade level could be avoided by instrumenting the share of girls at the grade level by the share of girls at the birth cohort level. We find that this approach would generally continue to yield spurious gender peer effects.

[6]Hoxby (2000) does not have the individual data necessary to assign students to birth cohorts (her data consist of grade level averages). Israeli primary schools participate only once every two years in the standardized tests used by Lavy and Schlosser (2011). As a result, test results are never available for both retained and non-retained students in a birth cohort.

[7]Hoxby (2000) and Lavy and Schlosser (2011) also find a positive effect of the share of girls on boys' academic achievement in subjects where girls do worse than boys on average. See Sacerdote (2011) for a review of the literature on skill spillovers in school.

we lower the academic thresholds for grade retention – which leads to lower rates of grade retention – and simulate data. The simulated data allow us to estimate gender peer effects at successively lower rates of grade retention. Our findings indicate that the birth cohort and the grade level approach may yield different patterns of gender peer effects at rates of grade retention around or below the OECD and EU average. However, in interpreting these findings it should be kept in mind that they are based on counterfactual experiments calibrated to Spanish data and therefore cannot be generalized.

The community of Madrid data contain information on the primary school students attended when they took the test but do not specify where students went to school previously. This is also true for the Texas primary school data used by Hoxby (2000) and the Israeli primary school data used by Lavy and Schlosser (2011). Student mobility across primary schools in Israel is substantially lower than in Texas however. Lavy and Schlosser calculate that 7.9 percent of students in Israel left their primary school between first and second grade in 2002, while Hanushek, Kain, and Rivkin (2004) report an annual rate of student mobility in Texas of 24 percent. In our data, student mobility in primary school appears to be comparatively low. We estimate that the share of second grade students in 2013 who did not attend the same primary school a year earlier is around 2.8 percent.

The remainder of the paper is structured as follows. We discuss the related literature next and then preview our theoretical findings on spurious gender peer effects when using a grade level approach. Section 2 develops our theoretical model of a school system with grade retention and uses it to discuss the grade level approach and the birth cohort approach to gender peer effects. Section 3 presents the data and section 4 our empirical results. Section 5 contains our counterfactual experiments. Section 6 concludes.

1.1 Related Literature

Our study is closely related to the work of Hoxby (2000) and Lavy and Schlosser (2011), who estimate gender peer effects using within-school variation in gender composition at the grade level in Texas and Israel respectively. Lavy and Schlosser estimate gender peer effects in primary, middle, and high school. In primary school, they find a statistically significant effect of the share of girls on girls' achievement in mathematics and an insignificant effect of girls on boys' mathematics achievement. In science and technology, they find a statistically significant gender peer effect of girls on girls' and on boys' achievement, while gender peer effects are statistically insignificant in Hebrew and English. In middle school, Lavy and Schlosser find statistically significant gender peer effects of girls on girls' achievement in mathematics and English but not in science and technology or Hebrew. Gender peer effects

of girls on boys' achievement are statistically insignificant in all subjects. In high school, Lavy and Schlosser find gender peer effects of girls on girls' as well as boys' achievement for a range of academic outcomes. Lavy and Schlosser also provide evidence on mechanisms. They find that a greater share of girls in a grade lowers levels of classroom disruption and violence, improves inter-student and student-teacher relationships, and lessens teacher fatigue. These effects are not driven by changes in the behavior of individual girls or boys but by the share of (better behaved) girls. Hoxby's study of gender peer effects in Texas primary schools finds statistically significant effects of girls on girls' achievement in mathematics and reading in grades three to six. Gender peer effects of girls on boys' achievement are also mostly significant in reading, with the exception of fourth grade. Gender peer effects of girls on boys' achievement in mathematics are statistically insignificant for fourth and fifth grade but positive and significant in third and sixth grade.

As Israel, Spain, and Texas differ in their socioeconomics (e.g. the role of woman in society), their school systems, and their standardized tests, estimates of gender peer effects need not be similar. The 95 percent confidence intervals for our preferred estimates of gender peer effects in Spanish primary schools are broadly similar to Lavy and Schlosser's (2011) for Israeli primary schools in mathematics, which is probably the most comparable subject. Hoxby's (2000) estimates for Texas are somewhat different even at similar stages in primary school. Hoxby's 95 percent confidence interval for the effect of a 10-percentage-point increase in the share of girls on girls' achievement in mathematics is from around 0 to around 12 percent of a standard deviation in fifth and sixth grade. The 95 percent confidence interval for Lavy and Schlosser's estimates of gender peer effects of girls on girls' achievement in mathematics is from around 0 to around 7 percent of a standard deviation, while it is from around -1 to around 4 percent of a standard deviation for our preferred estimates. The 95 percent confidence interval for Hoxby's estimates of the effect of a 10-percentage-point increase in the share of girls on boys' achievement in mathematics is from around -6 to around 5 percent of a standard deviation in fifth grade and from around 0 to around 16 percent of a standard deviation in sixth grade. The 95 percent confidence interval for Lavy and Schlosser's estimates of gender peer effects of girls on boys' achievement in mathematics is from around -1 percent to around 5 percent of a standard deviation, while it is from around 0 to around 5 percent of a standard deviation for our preferred estimates.

Our work is also related to Whitmore's (2005) study of gender peer effects in kindergarten and primary school grades one to three using differences in gender composition generated by the STAR project.[8] Her approach exploits differences in gender composition across classes

[8]The project randomly assigned kindergarten children in participating schools and students entering

within grades and schools. She finds that the effect of more girls in class on the average student's academic achievement depends on the grade. The effect is significantly positive in kindergarten and grade two, insignificant in grade one, and significantly negative in grade three.[9] When Whitmore allows for different effects on girls and boys, estimates become noisy and are mostly statistically insignificant (maybe because of the relatively small sample size compared to Hoxby or Lavy and Schlosser).[10]

1.2 Preview of Main Theoretical Findings

A main finding from our theoretical model of a school system with grade retention is that the grade level approach generally yields spurious gender peer effects even if grade level differences in gender composition were solely driven by natural fluctuations in the share of girls across birth cohorts. To understand this result consider the following scenario. Suppose that each incoming (first-grade) class in a primary school has the same size but that the share of girls is subject to exogenous shocks. Suppose also that individual students' skills are unaffected by the gender composition of their class (there are no true gender peer effects). Consider a class starting out with a larger share of girls than the average incoming class in the school. As this class proceeds from first grade to higher grades it will lose students who are retained and be joined by retained students. But the share of girls among all students in the class will tend to remain higher than in a class starting with the average gender composition.[11] Now consider the share of retained girls among all girls as the class proceeds from first grade to higher grades. As long as the gender of retained students is independent of the gender composition of the class they join, a larger-than-average share of girls in first grade leads to a smaller-than-average share of retained girls among all girls in the class in higher grades. As a result, exogenous shocks to the gender composition of birth cohorts translate into a negative association between the share of girls among all students and the share of retained girls among all girls in higher grades. This association is at the root of the spurious gender peer effects when using the grade level approach.

participating schools to classes (of different types).

[9] Graham, Imbens, and Ridder (2010) develop an approach to quantify the gains from reallocating individuals across social groups in the presence of spillovers and illustrate their approach by studying the effects of gender segregation on mathematics achievements in kindergarten using STAR data. Their analysis differs from Whitmore in that it only looks at gender peer effects in math; only in kindergarten; and it allows for nonlinear effects.

[10] There is also a literature using intention-to-treat approaches based on primary school enrollment rules to examine whether differences in initial maturity have long-lasting effects on academic achievement, see Bedard and Dhuey (2006).

[11] This will be the case even if the likelihood of retention differs for girls and boys as long as the gender of retained students is independent of the gender composition of the class they join. A sufficient condition for this to be the case is that gender composition is independently distributed across birth cohorts.

The direction of the spurious gender peer effects mainly depends on whether non-retained students do better or worse on average than students who were retained in the past. If (as in our data) non-retained students tend to do better, a class with a larger-than-average share of girls in first grade will tend to have girls who do better than average in higher grades because of a smaller-than-average share of worse-performing retained girls among all girls. As a result, exogenous shocks to the gender composition of birth cohorts translate into a positive association between the share of girls among all students and the academic achievement of girls in higher grades. This positive association between the share of girls and girls' academic achievement in higher grades produces a spurious positive effect of the share of girls on girls' achievement when using the grade level approach.[12] For boys, the argument is symmetric and implies a spurious positive gender peer effect of the share of boys in a grade on the academic achievement of boys. As the share of girls and boys in a grade sums to one, this translates into a spurious negative gender peer effect of the share of girls on the achievement of boys.

The grade level approach generally also yields spurious gender peer effects when there are exogenous shocks to the skills of girls and boys in a birth cohort. But such skill shocks do not – as a first analysis might suggest – necessarily translate into a spurious positive gender peer effect of girls on girls' academic achievement and of boys on boys' achievement. Instead, the direction of the spurious gender peer effects turns out to depend on whether grade repetition improves or worsens retained students' skills. To understand the main forces shaping the direction of the spurious gender peer effects induced by skill shocks, consider an incoming (first-grade) class where the share of girls is equal to the average in incoming classes but girls have better-than-average skills. As this class proceeds from first grade to higher grades, a greater-than-average share of girls will be promoted (not retained). Moreover, the skills of non-retained girls will be better than average. Hence, this class will end up with a larger share of girls among all students in higher grades and with better-than-average girls. This would tend to translate into a spurious positive effect of the share of girls on the academic achievement of girls when using the grade level approach.

However, there is a countervailing force operating in the class one year below the class that starts with better-than-average girls. As this class proceeds to higher grades, it receives retained girls from the class that starts with better-than-average girls. As the skills of these

[12]There turns out to be a countervailing force that could in principle dominate and reverse the sign of the spurious gender peer effect of girls on girls' achievement. This countervailing force emerges in the classes that receive retained students from the classes starting with a greater-than-average share of girls. We find that for this countervailing force to dominate, the grade retention policy has to be so stringent that more than half of the students repeat a grade. This scenario seems of little empirical relevance as rates of grade retention are below 50 percent in almost all countries (PISA, 2009a).

retained girls are above average and their number below average, the class one year below the class that starts with better-than-average girls ends up with better-than-average girls in higher grades and a smaller-than-average share of girls among all students. The dominant force in our theoretical model depends on whether grade repetition improves or worsens retained students' skills. If grade repetition improves retained students' skills, shocks to the skills of girls at the birth cohort level translate into a spurious positive gender peer effect of the share of girls in a grade on the academic achievement of girls. The effect of shocks to the skills of boys is symmetric and implies a spurious positive gender peer effect of the share of boys in a grade on the academic achievement of boys in the grade (or equivalently, a spurious negative effect of girls on boys' achievement).

Two important follow-up issues are whether spurious gender peer effects at the grade level can be avoided by controlling for grade retention at the individual level or by using an instrumental-variables approach, with the share of girls at the grade level instrumented by the share of girls at the birth cohort level. For the instrumental-variables approach to work, the share of girls at the birth cohort level should affect students' academic achievement solely through the share of girls at the grade level. We have already seen however that when there is grade retention and the average student retained in the past does not perform exactly as well as the average non-retained student in the same grade, the share of girls at the birth cohort level affects academic achievement at the grade level through the share of retained girls and boys among students of the same gender. This composition effect could be dealt with by controlling for grade retention at the individual level if within-school grade level differences in the share of girls were solely driven by natural fluctuations in the share of girls across birth cohorts. But this estimation strategy becomes invalid when grade level differences in the share of girls may also be driven by other factors, such as shocks to the academic skills of girls and boys in a birth cohort. The reason is that in this case, differences between the academic achievement of retained and non-retained students in a grade may reflect the same factors as differences in the academic achievement of students across school years.

2 A Theoretical Framework

To understand the consequences of grade retention for the estimation of gender peer effects within schools we develop a theoretical model that we can deal with analytically. The key feature of the model is that students with academic skills below a threshold are subject to grade retention. The model also features exogenous shocks to the gender composition of

birth cohorts, to the academic skills of girls and boys, and to the thresholds used for grade retention. We first use the model to examine if and why the grade level approach may give rise to gender peer effects although there are no true gender peer effects. We also use the model to illustrate the birth cohort approach to gender peer effects.

2.1 A School System with Grade Retention

Children in birth cohort t start primary school in year t. For the first L school years, children attend what we will call lower grades (LG) and all children in the same birth cohort are in the same classroom. At the end of the Lth year in primary school, some children move to what we will call high grade (HG) and some are retained for an extra year in LG. Children in HG take a standardized test at the end of the school year and then leave primary school.

Whether students are retained for an extra year in LG depends on how their individual academic skills a after L years in LG compare with their school's academic threshold for grade retention p. Students i of gender g in school s and birth cohort t move from LG to HG after L years if their academic skills at that point in time satisfy

$$a_{igs}^t \geq p_{gs}^t. \tag{1}$$

Students with skills below the academic threshold, $a_{igs}^t < p_{gs}^t$, are retained for one extra year in LG and therefore move to HG after $L+1$ years in LG. We sometimes refer to students in birth cohort t who move to HG in year $t+L$ and $t+L+1$ as students who enter HG on age and late respectively. The academic threshold p_{gs}^t may be subject to school, birth cohort, and gender specific shocks. (An alternative interpretation of shocks to p_{gs}^t are shocks to skills that are relevant for grade retention but irrelevant for the performance in the test students take in HG.)

Each year a continuum of children of measure one starts in each school. A share ϕ_s^t of the children entering school s in year t are girls and a share $1-\phi_s^t$ boys. The distribution of skills in birth cohort t in school s after L years in LG is taken to be uniform with density $1/2\theta$ and a gender, school, and birth cohort specific mean α_{gs}^t.[13]

Students' performance in the standardized test administered in HG depends on their skills when they reach HG and the skills they acquire in HG. The test performance of non-retained students from birth cohort t attending HG in the school year starting in τ is taken to reflect their academic skills $a_{igs}^t + w_{igs\tau}$ where $w_{igs\tau}$ refers to the skills that student i of gender g acquires by attending HG in the school year starting in τ. We assume that the $w_{igs\tau}$ of different students are obtained as independent draws from a distribution with a constant

[13] In our simulated models in section 5 we can consider a wider range of distributions.

variance and a mean $\omega_{gs\tau}$ that may depend on gender, school, and the school year. The test performance of students reaching HG after being retained for an extra year in LG is taken to reflect their academic skills $a^t_{igs} + w_{igs\tau} + \delta^t_{gs}$ where δ^t_{gs} captures a gender, school, and birth cohort specific change in skills associated with grade repetition. This change in skills may be positive or negative. If $\delta^t_{gs} > 0$, students who are retained for an extra year in LG accumulate additional skills and therefore do better in the standardized test than they would have done had they not been retained. Figure 1 summarizes the model.

Gender peer effects could end up affecting the academic skills of HG students and hence their test results in three main ways in the model. First, the share of girls among a student's peers during the first L years of LG could affect the student's academic skills a^t_{igs} at the end of the Lth school year. This can be captured by allowing average academic skills α^t_{gs} to depend on the share of girls in LG. Second, the share of girls among a student's peers could affect the threshold p^t_{gs} used to determine who is promoted to HG after L years in LG. Third, the academic skills $w_{igs\tau}$ a student accumulates in HG could depend on the share of girls among the student's HG peers. This can be captured by allowing average academic skills accumulated in HG $\omega_{gs\tau}$ to depend on the share of girls in HG.[14]

The rule for grade retention in (1) combined with the distribution of academic skills after L years in LG implies that the share of students of gender g in birth cohort t who reach HG in year $t + L$ (without being retained in LG) is

$$\lambda^t_{gs} = \frac{1}{2\theta}(\alpha^t_{gs} + \theta - p^t_{gs}) \tag{2}$$

We assume throughout that some but not all students are retained for an extra year in LG in each school, $0 < \lambda^t_{gs} < 1$, which amounts to the parameter restriction

$$-\theta < \alpha^t_{gs} - p^t_{gs} < \theta. \tag{3}$$

The average test performance of HG students from birth cohort t who are not retained in LG and reach HG in year $\tau = t + L$ is

$$E(test^t_{igs} \,|non\text{-}retained) = E(a^t_{igs} \,|a^t_{igs} \geq p^t_{gs}) + \omega_{gs,t+L} = \frac{\alpha^t_{gs} + \theta + p^t_{gs}}{2} + \omega_{gs,t+L} \tag{4}$$

where $E(a \,|a \geq p)$ denotes the average skills after L years in LG of students who are not retained and ω denotes the average skills these students accumulate in HG in year $t + L$. The average test performance of HG students from birth cohort t who are retained for an

[14] We could also allow for gender peer effects on the change in skills associated with grade repetition. This complicates the model without generating additional insights as far as we can see.

extra year in LG and reach HG in year $\tau = t + L + 1$ is

$$
\begin{aligned}
(5) \qquad E(test^t_{igs} \,|retained) &= E(a^t_{igs} \,|a^t_{igs} \leq p^t_{gs}) + \delta^t_{gs} + \omega_{gs,t+L+1} \\
&= \frac{1}{2}(\alpha^t_{gs} - \theta + p^t_{gs}) + \delta^t_{gs} + \omega_{gs,t+L+1}
\end{aligned}
$$

where $E(a \,|a < p)$ denotes the average skills after L years in LG of students who are retained, δ the change in skills associated with grade repetition, and ω the average skills these students accumulate in HG in year $t + L + 1$.

The average test performance of girls and boys in HG can be derived by combining (4)-(5) and the shares of non-retained girls and boys among HG students of the same gender. The share of non-retained girls among girls in HG in school s in the school year starting in τ depends on the share of girls in birth cohorts $\tau - L$ and $\tau - L - 1$ as well as the share of non-retained girls in these birth cohorts

$$
(6) \qquad \mu_{fs\tau} = \frac{\phi_s^{\tau-L}\lambda_{fs}^{\tau-L}}{\phi_s^{\tau-L}\lambda_{fs}^{\tau-L} + \phi_s^{\tau-L-1}(1 - \lambda_{fs}^{\tau-L-1})}
$$

where f stands for female, λ^t_{fs} is the share of non-retained girls among all girls in birth cohort t, and ϕ^t_s is the share of girls among all students in birth cohort t. The share of non-retained boys among HG boys can be obtained analogously

$$
(7) \qquad \mu_{ms\tau} = \frac{(1 - \phi_s^{\tau-L})\lambda_{ms}^{\tau-L}}{(1 - \phi_s^{\tau-L})\lambda_{ms}^{\tau-L} + (1 - \phi_s^{\tau-L-1})(1 - \lambda_{ms}^{\tau-L-1})}
$$

where m stands for male. The average test performance of HG students of gender g in school s and school year τ can now be obtained by combining (4)-(7)

$$
(8) \qquad test_{gs\tau} = \mu_{gs\tau} E(test^{\tau-L}_{igs} \,|non\text{-}retained) + (1 - \mu_{gs\tau}) E(test^{\tau-L-1}_{igs} \,|retained)\,.
$$

2.2 The Grade Level Approach to Gender Peer Effects

Could grade retention of academically weak students lead us to conclude that there are gender peer effects within schools although there are none? And in what direction might spurious gender peer effects go? To address these questions, we assume that there are no true gender peer effects within schools and ask what we would conclude if we assessed the strength of gender peer effects using a grade level approach.[15]

Suppose we have data on the test performance of HG students and the share of girls in HG for a large number of schools in school years τ and $\tau - 1$. The grade level approach would assess the strength of gender peer effects by regressing the test performance of HG girls or

[15] That is, we assume that α^t_{gs}, p^t_{gs}, and $\omega_{\tau gs}$ do not depend on the share of girls in LG or HG.

boys on the share of girls in HG and school fixed effects. This is equivalent to assessing the strength of gender peer effects by regressing the change in the average test performance of HG girls or boys between the school years starting in τ and $\tau-1$, $test_{gs\tau} - test_{gs,\tau-1}$, on the corresponding changes in the share of HG girls, $girlsh_{s\tau} - girlsh_{s,\tau-1}$. We have already derived $test_{gs\tau}$ in (8). The share of HG girls can be obtained by combining the share of girls and boys in birth cohorts $\tau-L$ and $\tau-L-1$ who are not retained in LG with the share of girls in these birth cohorts

$$(9) \qquad girlsh_{s\tau} = \frac{\phi_s^{\tau-L}\lambda_{fs}^{\tau-L} + \phi_s^{\tau-L-1}(1-\lambda_{fs}^{\tau-L-1})}{\phi_s^{\tau-L}\lambda_{fs}^{\tau-L} + \phi_s^{\tau-L-1}(1-\lambda_{fs}^{\tau-L-1}) + (1-\phi_s^{\tau-L})\lambda_{ms}^{\tau-L} + (1-\phi_s^{\tau-L-1})(1-\lambda_{ms}^{\tau-L-1})}.$$

We find that such a grade level approach generally yields spurious gender peer effects and that the direction of these effects depends on the impact of grade repetition on retained students' skills and the sources of shocks to the share of girls at the grade level. The possible sources of shocks in our model are shocks to the share of girls at the birth cohort level (ϕ_s^t), shocks to the academic skills of girls and boys at the birth cohort level ($\alpha_{fs}^t, \alpha_{ms}^t$), and shocks to the academic thresholds for grade retention (p_{fs}^t, p_{ms}^t). We examine the consequences of these three types of shocks in turn. To simplify the analysis somewhat, we assume throughout this section that the change in skills associated with grade repetition and mean skills accumulated in HG are constant, $\delta_{gs}^t = \delta$ and $\omega_{gs\tau} = \omega$.

2.2.1 Shocks to the Share of Girls at the Birth Cohort Level

Suppose that academic skills and the thresholds for grade retention are the same across schools, birth cohorts, and gender, $\alpha_{fs}^t = \alpha_{ms}^t = \alpha$ and $p_{fs}^t = p_{ms}^t = p$. The only shocks are i.i.d. shocks η_s^t with mean zero to the share of girls in a birth cohort

$$(10) \qquad \phi_s^t = 1/2 + \eta_s^t$$

where $Var(\eta_s^t) = Var(\eta) > 0$. What would we obtain when estimating gender peer effects in HG using the grade level approach? Linearizing the share of girls in HG in (9) yields that the grade level approach would yield a spurious positive gender peer effect of girls on girls' academic achievement and a spurious negative gender peer effect of girls on boys' achievement if

$$(11) \qquad (2\lambda - 1)(\theta - \delta) > 0$$

where $\theta - \delta$ is the difference between the average academic skills in HG of non-retained students and students retained in the past, see (4) and (5), and λ is the share of non-retained

students evaluated at α and p.[16] If $(2\lambda-1)(\theta-\delta)<0$, the grade level approach would yield a spurious negative gender peer effect of girls on girls' academic achievement and a spurious positive gender peer effect of girls on boys' achievement. Only if $(2\lambda-1)(\theta-\delta)=0$ would the grade level approach not yield spurious gender peer effects.

Hence, the grade level approach would indicate a spurious positive gender peer effect of girls on girls' academic achievement and a spurious negative gender peer effect of girls on boys' achievement if the school system's retention rate $1-\lambda$ is not too high, $1-\lambda<1/2$, and non-retained students do better on average in HG than students who were retained in the past, $\theta>\delta$. To get an intuitive understanding of these spurious gender peer effects, it is useful to think through the consequences of a small positive shock η to the share of girls in birth cohort t for the share of girls and their academic skills in HG in the school years starting in $t+L$ and $t+L+1$. Non-retained students from birth cohort t reach HG in $t+L$ and retained students in $t+L+1$. Making use of (9), the shock to the share of girls among HG students in $t+L$ and $t+L+1$ can be calculated as

$$\widehat{girlsh}_{s,t+L}=\lambda\eta \quad \text{and} \quad \widehat{girlsh}_{s,t+L+1}=(1-\lambda)\eta \tag{12}$$

where $\widehat{x}$ denotes deviations of x from its value in the absence of shocks. These expressions reflect that a share λ of the η extra girls in birth cohort t reach HG in year $t+L$ while a share $1-\lambda$ are retained for an additional year in LG and therefore only reach HG in year $t+L+1$.

The extra girls in birth cohort t who reach HG in year $t+L$ were not retained in the past and the increase in the share of girls among HG students in the school year starting in $t+L$ therefore goes together with an increase in the share of non-retained girls among HG girls. Using (6), the shock to the share of non-retained girls among HG girls can be calculated as

$$\widehat{\mu}_{fs,t+L}=2\lambda(1-\lambda)\eta. \tag{13}$$

The greater share of non-retained girls among HG girls translates into better average academic skills (and test performance) of HG girls in $t+L$ if non-retained students do better on average than students who were retained in the past. The shock to HG girls' average test performance in the school year starting in $t+L$ can be calculated using (8) and (13)

$$\widehat{test}_{gs,t+L}=\widehat{\mu}_{fs,t+L}(\theta-\delta)\eta=2\lambda(1-\lambda)(\theta-\delta)\eta. \tag{14}$$

Hence, if non-retained HG students have better academic skills on average than students retained in the past, $\theta>\delta$, a positive shock η to the share of girls in birth cohort t translates

[16] See the appendix for a proof of this result.

into a positive shock to the average test performance of HG girls in the school year starting in $t + L$.

But it turns out that $\theta > \delta$ also implies that the positive shock η to the share of girls in birth cohort t translates into a negative shock to the average test performance of HG girls in the school year starting in $t + L + 1$. To see this recall that the extra girls in birth cohort t who reach HG in year $t + L + 1$ were all retained in the past. Hence, the share of retained girls among HG girls increases and this worsens the average academic skills (and test performance) of HG girls if non-retained students do better on average than students who were retained in the past. The shock to the average test performance of HG girls can be calculated using (8)

$$\widehat{test}_{gs,t+L+1} = -\widehat{\mu}_{fs,t+L+1}\,(\theta - \delta)\,\eta = -2\lambda(1-\lambda)\,(\theta - \delta)\,\eta. \tag{15}$$

Combining (12) and (14)-(15) yields that if $\theta > \delta$, a positive shock η to the share of girls in birth cohort t leads to a positive grade level association between the average test performance of girls and the share of girls in HG in the school year starting in $t + L$ but a negative grade level association in the school year starting in $t + L + 1$. The grade level approach to gender peer effects ends up averaging these two associations and the sign of the spurious gender peer effect is determined by whether the positive or the negative association dominates.

To see this note that students in HG in any school year τ come either from birth cohort $\tau - L$ (non-repeaters) or birth cohort $\tau - L - 1$ (repeaters). Let η' be the shock to gender composition in birth cohort $\tau - L$ and η'' the independent shock to gender composition in birth cohort $\tau - L - 1$. It follows from (12) that the implied shock to the share of girls in HG in the school year starting in τ is $\lambda\eta' + (1-\lambda)\eta''$. The implied shock to the average test performance of HG girls is $2\lambda(1-\lambda)\,(\theta - \delta)\,(\eta' - \eta'')$ and can be obtained from (14)-(15). As the grade level approach assesses gender peer effects in academic achievement by regressing within-school changes in the test performance of HG girls on the corresponding changes in the share of girls across schools, the sign of the gender peer effect is determined by the sign of

$$E(\widehat{girlsh}_{s\tau}\widehat{test}_{fs\tau}\Big|\,\tau) = 2\lambda(1-\lambda)(2\lambda-1)\,(\theta-\delta)\,Var(\eta) \tag{16}$$

where we used $E(\eta'\eta'') = 0$. This implies that a grade level regression of changes in HG girls' test performance on changes in the share of HG girls yields a strictly positive effect if and only if (11) holds (we maintain throughout that $0 < \lambda < 1$). Intuitively, in this case, the grade level association between changes in the share of girls and their academic achievement in HG induced by non-retained girls is positive and dominates the negative

association induced by retained girls.

An analogous argument for the academic achievement of boys yields that if $(2\lambda - 1)(\theta - \delta) > 0$, a regression of within-school changes in HG boys' test performance on the corresponding changes in the share of HG boys yields a strictly positive effect. As the share of girls and boys in a grade sums to one, it follows that a regression of changes in HG boys' test performance on changes in the share of HG girls yields a strictly negative effect in this case.

2.2.2 Shocks to the Skills of Girls and Boys at the Birth Cohort Level

A second source of shocks to the share of girls at the grade level are shocks to the skills of girls and boys in a birth cohort. To understand the effect of these shocks suppose that the average skills of girls and boys are independent across schools and birth cohorts but possibly correlated within schools and birth cohorts

$$\alpha_{fs}^t = \alpha + \varepsilon_{fs}^t \quad \text{and} \quad \alpha_{ms}^t = \alpha + \varepsilon_{ms}^t \tag{17}$$

where $\varepsilon_{fs}^t, \varepsilon_{ms}^t$ are shocks with mean zero with $Var(\varepsilon_{fs}^t) = Var(\varepsilon_{ms}^t) = Var(\varepsilon) > 0$ and $Correl(\varepsilon_{fs}^t, \varepsilon_{ms}^t) = \rho_\varepsilon$. Linearizing the share of girls in HG in (9) yields that shocks to academic skills do not translate into shocks to the share of girls in HG if $\rho_\varepsilon = 1$. But if the correlation between the shocks to the average skills of boys and girls is less than perfect, $\rho_\varepsilon < 1$, skill shocks affect the share of girls in HG. In this case the grade level approach to gender peer effects would lead to a spurious positive gender peer effect of girls on girls' academic achievement and a spurious negative gender peer effect of girls on boys' achievement if

$$(2\lambda - 1)\delta > 0.^{17} \tag{18}$$

If $(2\lambda - 1)\delta < 0$, the grade level approach would yield a spurious negative gender peer effect of girls on girls' academic achievement and a spurious positive gender peer effect of girls on boys' achievement. Only if $(2\lambda - 1)\delta = 0$ would the grade level approach not yield spurious gender peer effects.

Hence, the grade level approach would indicate a spurious positive gender peer effect of girls on girls' academic achievement and a spurious negative gender peer effect of girls on boys' achievement if the retention rate $1 - \lambda$ is not too high, $1 - \lambda < 1/2$, and if grade repetition improves retained students' skills, $\delta > 0$. To get an intuitive understanding of these spurious gender peer effects, consider the implications of a small positive shock ε to the average skills of girls in birth cohort t for the share of HG girls and their academic skills

[17] See the appendix for a proof of this result.

in the school years starting in $t+L$ and $t+L+1$. As the positive shock to the skills of girls implies that fewer girls in birth cohort t are retained, it translates into a positive shock to the share of girls in HG in school year $t+L$ and a negative shock of the same size to the share of girls in HG in $t+L+1$,

$$\widehat{girlsh}_{s,t+L} = \frac{1}{8\theta}\varepsilon \quad \text{and} \quad \widehat{girlsh}_{s,t+L+1} = -\frac{1}{8\theta}\varepsilon \tag{19}$$

where we made use of (2) and (9). The positive shock to the skills of girls in birth cohort t also affects the average academic skills (and test performance) of HG girls. Making use of (8) yields that the average test performance of HG girls in the school years starting in $t+L$ and $t+L+1$ are shocked by

$$\widehat{test}_{fs,t+L} = \lambda\frac{1}{2}\varepsilon + (1-\lambda)\frac{1}{2}\left(1-\frac{\delta}{\theta}\right)\varepsilon \tag{20}$$

and

$$\widehat{test}_{fs,t+L+1} = (1-\lambda)\frac{1}{2}\varepsilon + \lambda\frac{1}{2}\left(1-\frac{\delta}{\theta}\right)\varepsilon. \tag{21}$$

These expressions capture two types of effects. An effect on test performance holding the share of non-retained girls constant and a composition effect that arises through a change in the share of non-retained girls. The effect holding the share of non-retained girls constant is captured by the first terms on the right-hand side of (20)-(21). These terms are obtained as the improvement in the average test performance of girls from birth cohort t who reach HG in the school years $t+L$ and $t+L+1$ multiplied by the weight of students from birth cohort t in these two school years. It is interesting to note that this effect on test performance is stronger in the school year where most girls in birth cohort t reach HG, which is school year $\tau+L$ as long as $\lambda > 1/2$. The second terms in (20)-(21) capture a composition effect as the positive skill shock increases the share of non-retained girls among HG girls in $t+L$ and $t+L+1$. If non-retained HG girls do better on average than girls retained in the past, $\theta > \delta$, this composition effect increase test performance in HG in both school years. This effect is weaker in the school year where most girls in birth cohort t reach HG.

Combining (19) and (20)-(21) yields that a positive shock ε to the average skills of girls in birth cohort t leads to a positive grade level association between the average test performance of girls and the share of girls in HG in the school year starting in $t+L$ but a negative grade level association in the school year starting in $t+L+1$ as long as $\delta \leq \theta$. When $\delta = \theta$ the positive association is stronger than the negative association but this changes as δ falls. When $\delta = 0$, the two associations exactly offset each other and the positive association is weaker than the negative association when $\delta < 0$. The grade level approach to gender peer

effects ends up averaging these two associations and the sign of spurious gender peer effects is determined by whether the positive or the negative association dominates.

To see this recall that students in HG in any school year τ come either from birth cohort $\tau-L$ or birth cohort $\tau-L-1$. Let ε' be the shock to the academic skills of girls in birth cohort $\tau-L$ and ε'' the independent shock to the academic skills of girls in birth cohort $\tau-L-1$. It follows from (19) that the implied shock to the share of girls in HG in the school year starting in τ is $\varepsilon'/8\theta-\varepsilon''/8\theta$. The implied shock to the average academic skills (and test performance) of HG girls can be calculated using (20)-(21) as $(1/2-(1-\lambda)\delta/2\theta)\varepsilon'+(1/2-\lambda\delta/2\theta)\varepsilon''$. As the grade level approach assesses gender peer effects of girls on girls' academic achievement by regressing within-school changes in the test performance of HG girls on the corresponding changes in the share of girls across schools, the sign of gender peer effects is determined by the sign of

$$E(\widehat{girlsh}_{s\tau}\widehat{test}_{fs\tau}\Big|\,\tau)=\frac{(2\lambda-1)\delta}{16\theta^2}Var(\varepsilon) \tag{22}$$

where we used that $E(\varepsilon'\varepsilon'')=0$. This implies that the grade level approach yields a strictly positive effect of girls on girls' academic achievement if the school system's retention rate $1-\lambda$ is not too high, $1-\lambda<1/2$, and $\delta>0$. In this case the positive grade level association between changes in the share of girls in HG and their academic achievement induced by non-retained girls dominates the negative association induced by retained girls. If $\delta\leq 0$, the positive association induced by non-retained girls is either exactly offset by the negative association induced by retained girls ($\delta=0$) or is dominated by the negative association induced by retained girls ($\delta<0$).

An analogous argument for the academic achievement of boys yields that if $(2\lambda-1)\delta>0$, a regression of within-school changes in HG boys' test performance on the corresponding changes in the share of HG boys yields a strictly positive effect. As the share of girls and boys in a grade sums to one, it follows that a regression of changes in HG boys' test performance on changes in the share of HG girls yields a strictly negative effect in this case.

2.2.3 Shocks to the Academic Thresholds for Grade Retention

A third source of shocks to the share of girls at the grade level are shocks to the academic thresholds for grade retention.[18] These shocks can be analyzed in a similar way as skill shocks. Suppose that the academic thresholds applied to girls and boys are subject to shocks that are independent across schools and birth cohorts but possibly correlated within

[18] As already mentioned an alternative interpretation of shocks to these academic thresholds are shocks to skills that are relevant for grade retention but irrelevant for the performance in the standardized test.

schools and birth cohorts

$$p^t_{fs} = p + \nu^t_{fs} \quad \text{and} \quad p^t_{ms} = p + \nu^t_{ms} \tag{23}$$

where ν^t_{fs}, ν^t_{ms} are shocks with mean zero with $Var(\nu^t_{fs}) = Var(\nu^t_{ms}) = Var(\nu) > 0$ and $Correl(\nu^t_{fs}, \nu^t_{ms}) = \rho_\nu$. Suppose also that the average skills of boys and girls are identical and the same across schools, birth cohorts, and gender, $\alpha^t_{fs} = \alpha^t_{ms} = \alpha$, and that there is the same share of boys and girls in each birth cohort, $\phi^t_s = 1/2$.

Linearizing the share of girls in HG in (9) yields that shocks to the academic thresholds for grade retention do not translate into shocks to the share of girls in HG if the shocks to the academic thresholds applied to girls and boys are perfectly correlated. If the correlation is less than perfect, $\rho_\nu < 1$, the grade level approach to gender peer effects would lead to a spurious negative gender peer effect of girls on girls' academic achievement and a spurious positive gender peer effect of girls on boys' achievement if

$$(2\lambda - 1)\left(1 - \frac{\delta}{2\theta}\right) > 0.^{19} \tag{24}$$

If $(2\lambda - 1)(1 - \delta/2\theta) < 0$, the grade level approach would yield a spurious positive gender peer effect of girls on girls' academic achievement and a spurious negative gender peer effect of girls on boys' achievement. Only if $(2\lambda - 1)(1 - \delta/2\theta) = 0$ would there be no spurious gender peer effects.

The condition in (24) can be derived analogously to (18). Intuitively, a positive shock to the academic threshold used for girls in birth cohort t leads to a negative grade level association between the average academic skills and the share of HG girls in the school year starting in $t + L$ but a positive grade level association in the school year starting in $t + L + 1$. When (24) holds, the negative association dominates and produces a spurious negative gender peer effect of girls on girls' academic achievement when using the grade level approach. The argument for the effect on boys' academic achievement is analogous.

2.2.4 A Grade Level Instrumental-Variables Approach?

It is natural to wonder whether spurious gender peer effects at the grade level could be avoided by using an instrumental-variables approach, with the share of girls in HG in the school year starting in τ instrumented by the share of girls in birth cohort $\tau - L$. This approach eliminates the spurious gender peer effects due to shocks to academic skills at the birth cohort level and academic thresholds used for grade retention described in sections 2.2.2 and 2.2.3. But the approach would continue to yield spurious gender peer effects. When

[19] See the appendix for a proof of this result.

there is grade retention, birth cohorts with relatively more girls end up having a lower share of retained girls among all girls in HG and a greater share of retained boys among all boys in HG as described in section 2.2.1. This composition effect continues to translate into spurious gender peer effects when using an instrumental-variables approach as long as the average student retained in the past does not perform exactly as well as the average non-retained student in HG. The composition effect could be dealt with by controlling for grade retention at the individual level if grade level differences in the share of girls were solely driven by exogenous shocks to the share of girls across birth cohorts. But this estimation strategy becomes invalid when grade level differences in the share of girls are also driven by shocks to academic skills or retention thresholds, as differences between the academic performance of retained and non-retained students in HG reflect some of the same factors as differences in the academic performance of students across school years in this case.

2.3 The Birth Cohort Approach to Gender Peer Effects

The birth cohort approach to gender peer effects examines whether girls or boys in a birth cohort with a greater share of girls do better academically than students of the same gender in other birth cohorts in the same school. We now illustrate the approach in our model.

As already mentioned when we set up the model, gender peer effects could affect academic achievement in three main ways. First, there could be gender peer effects in the accumulation of skills in lower grades (LG). To capture this effect we allow mean skills after L years in LG to depend on the share of girls in birth cohort t

$$\alpha_{gs}^{t} = \widetilde{\alpha}_{gs}^{\,t} + \pi_{g}^{\alpha}\phi_{s}^{t} \tag{25}$$

where π_g^α is the strength of the gender peer effect on the academic skills of students of gender g and $\widetilde{\alpha}_{gs}^{\,t}$ captures exogenous shocks to skills.[20] Gender peer effects could also affect the academic thresholds used to determine who is retained for an extra year in LG. We therefore also allow the thresholds for grade retention to depend on the share of girls in birth cohort t

$$p_{gs}^{t} = \widetilde{p}_{gs}^{\,t} + \pi_{g}^{p}\phi_{s}^{t} \tag{26}$$

where π_g^p is the strength of the gender peer effect on the threshold for grade retention applied to students of gender g and $\widetilde{p}_{gs}^{\,t}$ captures exogenous shocks. Moreover, there could be gender

[20] In the model, the share of girls in a student's birth cohort is identical to the share of girls in the student's classroom and grade during the first $L-1$ years of LG (we assumed one classroom per grade). In year L of LG, the share of girls in a student's classroom and grade starts differing from the share of girls in the student's birth cohort because of retained students from the birth cohort that is one year older. As a result, the formulation in (25) is an approximation to classroom or grade level gender peer effects in LG. It would be straightforward to extend the model to capture classroom or grade level gender peer effects in LG precisely but this would not generate additional insights as far as we can see.

peer effects in high grade (HG) skill accumulation. To capture this effect we allow mean academic skills accumulated in HG in school year τ to depend on the share of girls in HG defined in (9)

$$\omega_{gs\tau} = \widetilde{\omega}_{gs\tau} + \pi_g^{\omega} girlsh_{s\tau} \tag{27}$$

where π_g^{ω} is the strength of the gender peer effect and $\widetilde{\omega}_{gs\tau}$ captures exogenous shocks.

The birth cohort approach to gender peer effects within schools assesses the strength of gender peer effects by regressing the test performance of girls or boys in a birth cohort on the share of girls in the birth cohort and school fixed effects. This is equivalent to assessing the strength of gender peer effects by regressing the change in the average test performance of girls or boys between birth cohorts t and $t-1$, $test_{gs}^t - test_{gs}^{t-1}$, on the change in the share of girls between these birth cohorts, $\phi_s^t - \phi_s^{t-1}$. The average test performance of students in birth cohort t is the weighted average of the test performance of non-retained and retained students in the birth cohort

$$test_{gs}^t = \lambda_{gs}^t E(test_{igs}^t \,|non\text{-}retained) + (1-\lambda_{gs}^t) E(test_{igs}^t \,|retained) \tag{28}$$

where λ_{gs}^t is the share of non-retained students of gender g in birth cohort t and school s among all students of the same gender. Using (2), (4), and (5) this simplifies to

$$test_{gs}^t = \alpha_{gs}^t + \left(\lambda_{gs}^t \omega_{gs,t+L} + (1-\lambda_{gs}^t)\omega_{gs,t+L+1}\right) + \delta_{gs}^t (1-\lambda_{gs}^t). \tag{29}$$

The average test performance of students in a birth cohort is therefore the sum of three terms. The first term captures average academic skills of students in the birth cohort after L years in LG and the second term the academic skills these students accumulate in HG. The third term captures that retained students, who represent a share $1-\lambda_{gs}^t$ of the birth cohort, experience a change in academic skills during the extra year in LG.

Using (29) and (25)-(27), the strength of gender peer effects when using the birth cohort approach can be calculated as

$$\beta_g^B = \pi_g^{\alpha} + \pi_g^{\omega}\left(\lambda^2 + (1-\lambda)^2 + \lambda \frac{\partial(\lambda_{fs} - \lambda_{ms})}{\partial \phi}\right) - \delta \frac{\partial \lambda_{gs}}{\partial \phi} \tag{30}$$

where we have linearized (29) around $\widetilde{\omega}_{gs\tau} = \widetilde{\omega}$, $\widetilde{p}_{gs}^{\,t} = \widetilde{p}$, $\widetilde{\alpha}_{gs}^{\,t} = \widetilde{\alpha}$, $\delta_{gs}^t = \delta$ as well as $\phi_s^t = 1/2$, and $\lambda_{gs} = (\widetilde{\alpha} + \pi_g^{\alpha}\phi + \theta - \widetilde{p} + \pi_g^{p}\phi)/2\theta$ combines (2) and (25)-(26). The three terms in (30) correspond to the marginal effects of the share of girls in a birth cohort on the three terms in (29). The first term captures the strength of gender peer effects during the first L years in LG. The second term captures the expected strength of gender peer effects on the average girl or boy in HG. It is the product of the strength of gender peer effects in HG π_g^{ω} and a

term (in parentheses) that captures the expected increase in the share of girls the average student in a birth cohort with more girls will be exposed to in HG.[21]

To understand the third term on the right-hand side of (30) it is useful to focus on the case where gender peer effects arise solely through the academic thresholds used for grade retention ($\pi_g^p \neq 0$ but $\pi_g^\alpha = \pi_g^\omega = 0$ in (25)-(27)). Suppose for example that more boys are retained in birth cohorts with a greater share of girls ($\partial\lambda_{ms}/\partial\phi < 0$). As grade repetition changes skills by δ, the average boy in a birth cohort with more girls would then do better in HG than the average boy in a birth cohort with a balanced gender composition if $\delta > 0$.

A key feature of the birth cohort approach to gender peer effects that follows from (30) is that it only yields gender peer effects if gender composition affects the academic skills of students or the academic thresholds used for grade retention. To see this, note that $\pi_g^\alpha = \pi_g^\omega = \pi_g^p = 0$ combined with (2) and (25)-(27) implies $\beta_g^B = 0$.

3 Background and Data

The community of Madrid, one of Spain's largest and wealthiest regions, has been administering a standardized test to students in sixth grade (the last grade of primary school) since the school year starting in 2004.[22] Since 2008, test results come accompanied by a range of student characteristics, that include, for example, birth year and month, education and occupation of students' parents, and country of birth. These data plus the name of the school sixth graders attended have been made available to us for three school years (the school years starting in 2008, 2009, and 2010). The data cover around 50,000 students per school year in around 1150 primary schools. See appendix table 1 for summary statistics.

The Spanish primary school enrollment rule is that children start school in the year they turn six years old and more than 99 percent of children follow this rule in Spain and the community of Madrid (about 0.5 percent of children start a year later and 0.5 percent a year earlier).[23] Spanish primary schools permit grade retention and rates of grade retention

[21]To see this, note that λ and $1-\lambda$ are the probabilities that a student is not retained and retained respectively. Moreover, if gender peer effects did not affect retention rates, (9) implies that a one-percentage-point increase in the share of girls in a birth cohort increases the share of girls in HG retained students will be exposed to by $1-\lambda$ percentage points and the share of girls in HG non-retained students will be exposed to by λ percentage points. Hence, the increase in the share of girls the "average" student in a birth cohort with a one-percentage-point greater share of girls will be exposed to in HG is $\lambda\lambda + (1-\lambda)(1-\lambda)$. The third term in the parentheses in (30) captures that the share of girls in a birth cohort may also affect the share of girls in HG through a differential gender peer effect of girls on the retention rates of girls and boys.

[22]The community of Madrid's population in 2013 was around 6.5 million and income per capita close to 30 000 euros (Eurostat Regional Yearbook, 2013).

[23]This is higher than in countries like Iceland and Norway, which are known for almost all children complying with the enrollment rule (e.g. Bedard and Dhuey, 2006). We thank the community of Madrid's department of education, youth, and sports for providing this information for the community of Madrid. For

in primary school in Spain and the community of Madrid are higher than the OECD or European Union average. The share of 15 year olds in 2009 who report having been retained at least once in primary school is 11 percent in Spain and 11.8 percent in the community of Madrid, while it averages to 8 percent across OECD countries and to 7 percent across EU countries (PISA, 2009a). The share of students in Spain and the community of Madrid who report having been retained twice or more during primary school is 0.5 percent, which is similar to the OECD and EU average. It is likely that many of these students were retained at least once outside of Spain, as the community of Madrid forbids that students are retained twice during primary school and the rest of Spain allows it only in very exceptional cases.[24]

Our data for sixth graders in the community of Madrid do not contain information on whether students have repeated a grade. We know students' birth years however and can therefore check whether students born in year $t-6$ took the test during the school year starting in year $t+5$, as they should have if they started primary school according to the enrollment rule and were not retained. We find that 14.2 percent of students took the test one year late and 0.5 percent two or more years late. Hence, the share of students who repeated twice or more in primary school according to PISA coincides with the share of students who took the test two or more years late in our data. To see whether the PISA statistic on the share of students who repeated once in primary school in the community of Madrid is consistent with the share of students who took the test one year late in our data, we need to take into account that around 0.5 percent of students start primary school one year later than specified by the enrollment rule.[25] These students end up taking the test at least one year late even if they are not retained in primary school. Moreover, PISA allows participating countries and regions to exclude special needs students, who represent 2.3 percent of the students in our data, and students who have language difficulties because they arrived only recently from abroad (PISA, 2009b). As most of the students excluded from PISA are likely to have repeated a grade (95 percent of special needs students in our data take the test at least one year late) the PISA statistic and our statistic appear to be consistent.

The standardized test administered to sixth graders in the community of Madrid has four components: mathematics, reading, dictation, and general knowledge. Each component is scored between 0 and 10 (this is the usual range for grades in Spain). We transform these raw scores into standard scores $a_{i\tau} = (z_{i\tau} - \mu_\tau)/\sigma_\tau$, where $z_{i\tau}$ is the raw score of test taker

Spain the data come from the national statistical institute's INEbase (2013).

[24]Throughout Spain it is not permitted to retain students more than once in the same grade. See Spanish laws 10/2002 and 2/2006 and decree 22/2007 of the community of Madrid.

[25]These children can be thought of as being retained in kindergarten.

i in the school year starting in year τ, and μ_τ and σ_τ are the mean and standard deviation of the raw scores. We report results on gender peer effects based on an average across all test components and on the mathematics component only. We also considered the reading component only and did not find evidence of gender peer effects.[26]

We implement the birth cohort approach to gender peer effects for students born in 1997 and 1998. According to the Spanish enrollment rule, these students should have started primary school in 2003 and 2004 respectively (we index birth cohorts by the year they should have started primary school). As our data cover sixth graders in the school years starting in 2008, 2009, and 2010, we observe students from the 2003 birth cohort if they started according to the enrollment rule and were not retained in any grade; if they took the test one year late; or if they took the test two years late. Students from the 2003 birth cohort are not in our data if they took the test one or more years early (during or before the school year starting in 2007) or three or more years late (during or after the school year starting in 2011). We observe students from the 2004 birth cohort if they started according to the enrollment rule and were not retained; if they took the test one year early; or if they took the test one year late. Students from the 2004 birth cohort are not in our data if they took the test two or more years early (during or before the school year starting in 2007) or two or more years late (during or after the school year starting in 2011). As only a small share of students enter primary school early or are retained more than once during primary school in the community of Madrid and Spain, we end up missing a small share of students from the 2003 and the 2004 birth cohorts due to data availability issues. We can assess how many students we miss from each birth cohort by calculating the share of sixth graders in the school years starting in 2008, 2009, and 2010 who took the test early or two years late. Around 0.5 percent of sixth graders took the test two or more years late in our data and 0.3 percent of sixth graders took the test one or more years early. As a result, we estimate that we miss around 0.4 percent of the students of the 2003 and 2004 birth cohorts because of data availability issues. We treat the 2003 and 2004 birth cohorts symmetrically and drop students from the 2004 birth cohort who took the test one year early (in the school year starting in 2008) because we cannot observe students from the 2003 birth cohort who took the test one year early (in the school year starting in 2007). We also drop students from the 2003 birth cohort who took the test two years late (in the school year starting in 2010) because we cannot observe students from the 2004 birth cohort who took the test two years late (in the school year starting in 2011). As we treat the 2003 and 2004 birth cohorts

[26]Interestingly, Lavy and Schlosser (2011) find that gender peer effects are statistically insignificant in Hebrew, which is probably closest to reading in our data.

symmetrically, we end up missing around 0.9 percent of the students from these birth cohorts because of data availability.

Students who were retained in sixth grade attend sixth grade twice. We want to focus on students' academic achievement at the end of primary schooling (sixth grade is the last grade of primary school in Spain) and therefore only include these students in our empirical analysis in their last year of primary schooling. Our data do not specify whether students repeated sixth grade however. We therefore proceed as follows. Based on information on fifth and sixth graders in the community of Madrid in the school years starting in 2009 and 2010 from the Spanish national statistical institute's INEbase (2013) database, we estimate that 2.5 percent of sixth graders were repeating sixth grade.[27] As there are approximately 50,000 students attending sixth grade in the school year starting in 2009 in our data, we estimate that around 1250 students are repeating sixth grade. These are students from the 2003 birth cohort who should have attended sixth grade for the first time one year earlier, in the school year starting in 2008. It is useful to denote the set of students from the 2003 birth cohort who were in sixth grade in the school year starting in 2008 by $S(2003,2008)$ and the set of students from the 2003 birth cohort who were in sixth grade in the school year starting in 2009 by $S(2003,2009)$. The first set contains students from the 2003 birth cohort who reached sixth grade on age while the second set contains students from the 2003 birth cohort who were attending sixth grade one year late. The approximately 1250 students from the 2003 birth cohort who were repeating sixth grade in the school year starting in 2009 should be in both sets. We therefore look for all pairs of observations (i, j) with i from $S(2003,2009)$ and j from $S(2003,2008)$ where i and j attend the same school; were born in the same month and country; arrived to Spain at the same age if they were born abroad; and also coincide in terms of the country of birth and the level of education of mothers.[28] This yields 1172 pairs, which is close to the 1250 pairs we expected based on the retention rates from INEbase (2013). We then proceed under the assumption that these are the students from the 2003 birth cohort who repeated sixth grade in the school year starting in 2009. We use the same approach to identify students from the 2004 birth cohort who were repeating sixth grade in the school year starting in 2010. In this case the approach yields 1246 matching pairs, again close to the 1250 pairs we expected.

[27] INEbase contains the share of fifth and sixth graders in the community of Madrid who are repeating fifth or sixth grade in these school years. This share has been very stable at around 2.5 percent since there are data (the school year starting in 2001). INEbase does not have information for sixth grade only.

[28] We focus on mothers because according to the Spanish national statistical institute's INEbase (2013) database, 90 percent of children who live with one parent only, live with their mother. Around 9 percent of the observations have missing values for the education level of mothers. In these cases we look for observations i and j that coincide in that the education of mothers is missing.

Student mobility across primary schools in the community of Madrid appears to be low compared to Texas and Israel. Only 1.9 percent of the students in the second grade of public primary schools in the school year starting in 2012 did not attend the same school a year earlier and the analogous statistics for grades three through six are very similar.[29] For privately managed schools there are no official data on student mobility. We therefore surveyed 224 privately managed primary schools (50 percent of the privately managed schools in our sample). We obtained responses from 198 schools. In these schools, 3.9 percent of the students in second grade in the school year starting in 2013 did not attend the same school a year earlier. As around 55 percent of the students in our data attend public schools, we estimate average annual mobility between first and second grade of primary school in the community of Madrid to be 2.8 percent.[30] For comparison, Lavy and Schlosser calculate that 7.9 percent of students left their primary school between first and second grade in Israel in 2002, while Hanushek, Kain, and Rivkin (2004) report an annual rate of student mobility in Texas of 24 percent.

4 Empirical Results

We now employ both the birth cohort approach and the grade level approach to estimate gender peer effects in our sample of Spanish primary schools.

4.1 Results Using the Birth Cohort Approach

We start by checking the balancedness of student characteristics with respect to the share of girls in the birth cohort. We have data on a range of student characteristics: the education levels of mothers and fathers (4 categories); the profession/occupation of mothers and fathers (8 categories); with whom students live (6 categories); when students started preschool/school (4 categories); where students and their parents were born; at what age students arrived in Spain if they were born abroad; and whether students have some disability.[31]

[29] They are between 1.7 and 2.2 percent. We are very grateful to the community of Madrid's department of education, youth, and sports for providing this information.

[30] The likely causes of low school mobility in Spain are low residential mobility and a school assignment system for public as well as publicly funded, privately managed schools that strongly incentivates parents to pick a neighborhood school (e.g. Caldera and Andrews, 2011; Calsamiglia and Guell, 2013).

[31] For around 13 percent of students, we do not have the education of fathers and for around 9 percent we do not have the education of mothers. For profession/occupation we do not have data for around 7 percent of fathers and 12 percent of mothers. For other characteristics, the data is almost complete (we miss data for at most 0.1 percent of students).

To check whether the characteristics of students are balanced with respect to the share of girls in their birth cohort, we use the following estimating equation

$$x_{igs}^{jt} = \gamma_{gs}^{j} + \gamma_{g}^{jt} + \gamma_{g}^{j} Birthcohort Girlshare_{s}^{t} + u_{igs}^{jt} \tag{31}$$

where x^j stands for student characteristic $j = 1, ..., J$; i refers to individuals, g to gender, s to schools, t to the 2003 and 2004 birth cohorts; and u_{igs}^{jt} is the residual. School level differences and gender specific trends in characteristic j for gender g are captured by γ_{gs}^{j} and γ_{g}^{jt} respectively. The parameter of interest is γ_{g}^{j} which captures whether student characteristic j varies with the share of girls in the birth cohort.

Table 1 contains our estimates of γ_{g}^{j} in (31) as well as robust standard errors clustered by school. We also report a statistic for the hypothesis that $\gamma_{g}^{1} = ... = \gamma_{g}^{J} = 0$ for girls as well as boys. The results in the table are for schools with at most two classes per grade as our main results are for these schools. The results in table 1 indicate that student characteristics are balanced with respect to the share of girls in a birth cohort.[32] Of the 74 characteristics we examine for boys and girls, only 4 yield statistically significant γ_{g}^{j} at the 5 or 10 percent level (the number one would expect based on type I error). Moreover, the hypothesis $\gamma_{g}^{1} = ... = \gamma_{g}^{J} = 0$ cannot be rejected at any conventional significance level (the p-values are 0.69 for girls and 0.64 for boys).

We also use a second approach to see whether students in a birth cohort with a greater share of girls may have characteristics that are associated with better academic performance. The starting point are separate regressions of girls' and boys' individual test scores on school and birth cohort fixed effects, student characteristics, and indicators for missing student characteristics.[33] We then use the results to obtain a predicted test score for each student. This predicted test score is our best estimate of how well a student does academically given his or her characteristics. We then use the predicted test score as the left-hand-side variable in (31) to check balancedness with respect to the share of girls in the birth cohort. The advantage of this approach is that it quantifies any non-balancedness of student characteristics in terms of test performance.

Table 2 summarizes our results on the link between the predicted test scores of students and the share of girls in their birth cohort. Panel A reports the results for schools with at

[32] We also checked for evidence of unbalancedness in missing data. To do so we replaced the left-hand side of (31) by an indicator variable that takes the value of one if and only if we do not have information on characteristic j for individual i. This did not yield evidence of unbalancedness in missing data.

[33] The regressions are $test_{igs}^{t} = \alpha_{gs} + \alpha_{tg} + \sum_{j} \alpha_{g}^{j} x_{igs}^{jt}(1 - D_{igs}^{jt}) + \sum_{j} \gamma_{g}^{j} D_{igs}^{jt} + v_{igs}^{t}$ where $test_{igs}^{t}$ refers to standard test scores, i to individuals, $t = 2003, 2004$ to birth cohorts, g to gender, s to schools, D_{igs}^{jt} is an indicator variable that takes the value of one if and only if we do not have information on the student's characteristic j, and v_{igs}^{jt} is the residual. Individual characteristics account for around 11 to 15 percent of the variance in test performance.

most two classes per grade (the sample used in table 1). Predicted test scores of students appear to be balanced for girls as well as boys. For boys, we obtain a point estimate of -0.007 with a clustered standard error of 0.034 while for girls we obtain a point estimate of 0.007 with a standard error of 0.036. The point estimates imply that a 10-percentage-point increase in the share of girls in a birth cohort is associated with an improvement in the predicted test performance of less than one-tenth of one percent of a standard deviation for girls and a worsening in the predicted test performance of less than one-tenth of one percent of a standard deviation for boys. Panels B and C report the coefficient estimates for the sample of schools with only one class per grade and for all schools respectively. These results also indicate that there is no statistically significant link between the share of girls in a birth cohort and the predicted test performance of students in the birth cohort.

Table 3 contains our estimates of gender peer effects using the birth cohort approach. All results are based on the estimating equation

$$test^t_{igs} = \alpha_{gs} + \alpha_{tg} + \alpha_g X^t_{igs} + \beta_g Birthcohort Girlshare^t_s + v^t_{igs} \tag{32}$$

where $test^t_{igs}$ refers to individual standard test scores, i to individuals, $t = 2003, 2004$ to birth cohorts, g to gender, and s to schools; X_i is a vector collecting the characteristics j of student i and indicators for missing characteristics (if any); and v^t_{igs} is the residual. The table reports the least squares estimates of β_g with robust standard errors clustered by school. We report results for the average test score and the test score in mathematics for all schools, schools with at most two classes per grade, and schools with one class per grade.

Table 3, panel A examines the evidence for gender peer effects when we include all 1155 schools in our analysis. Around 20 percent of these schools have more than three classes per grade (the maximum number of classes per grade is five).[34] As the share of girls in a birth cohort and school is a noisy measure of the share of girls in any specific classroom, this could lead to noisy estimates of gender peer effects if these effects are at the classroom level. Our estimates in table 3, panel A indicate statistically insignificant gender peer effects when we use the average test score to measure academic achievement. This continues to be the case when we control for all the individual characteristics in table 1 plus indicators for missing characteristics and also when we control for peer group characteristics at the birth cohort level. The peer group characteristics included are the share of mothers as well as fathers with each of the four education levels in table 1, the share of immigrants, and the number

[34]The median number of students in schools with at most two classes per grade is 36 and the maximum number of students 58. The median number of students in schools with more than two classes per grade is 73 and the maximum number 152.

of students. Only when we measure academic achievement with the mathematics test score is there some evidence that boys do better when there are more girls in their birth cohort.

Table 3, panel B examines the evidence for gender peer effects when we only include the 908 schools with less than two classes per grade in our analysis. On the one hand, dropping schools with more than two classes per grade reduces the sample size, which may make estimates noisier. On the other hand, there should be a closer link between the share of girls at the birth cohort level and the classroom level in schools with fewer classes per grade and this should make it easier to pick up gender peer effects at the classroom level (if any). Our estimates in table 3, panel B indicate statistically significant gender peer effects of girls on the academic achievement of boys but not girls. This is the case whether we measure academic achievement using the average test score or the mathematics test score, and the estimates are very similar when we control for individual characteristics and peer group characteristics. The point estimates indicate that a 10-percentage-point increase in the share of girls in a birth cohort improves the test performance of boys by around 2.5 percent of a standard deviation.

Table 3, panel C examines the evidence for gender peer effects when we only include the 331 schools with one class per grade in our analysis. This is a substantially smaller sample of schools, which could make estimates much noisier. On the other hand, the link between the share of girls at the birth cohort level and the classroom level is as close as can be. Our estimates continue to suggest statistically significant gender peer effects of girls on the academic achievement of boys but not girls for the average test score and the mathematics test score when we control for individual characteristics and peer group characteristics. The point estimates indicate that a 10-percentage-point increase in the share of girls in a birth cohort improves the average test score of boys by around 3 percent of a standard deviation and the mathematics test score by around 4 percent of a standard deviation.[35]

4.2 Results Using the Grade Level Approach

Table 4 reports our estimates of gender peer effects using the grade level approach. Results are based on the grade level estimating equation

$$test_{igs\tau} = \alpha_{gs} + \alpha_{g\tau} + \alpha X_{igs\tau} + \beta_g GradeGirlshare_{s\tau} + v_{igs\tau} \tag{33}$$

where all variables are defined analogously to (32) and $\tau = 2008, 2009$ refers to the school years starting in 2008 and 2009. Overall, the grade level estimates of gender peer effects

[35] We do not find statistically significant effects of the share of girls in a birth cohort on the probability of grade retention of individual students. The grade retention effect captured by the last term on the right-hand side of (30) does therefore not appear to be at work in our data.

point in a different direction than the birth cohort estimates in table 3. For the two larger samples in panels A and B, we obtain a statistically significant effect of the share of girls on the academic achievement of girls but not on the achievement of boys.[36]

The difference in the pattern of gender peer effects using the birth cohort approach in table 3 and the grade level approach in table 4 is consistent with our theoretical model if (i) grade repetition improves retained students' skills but students retained in the past still tend to do worse on average than non-retained students in the same grade and (ii) grade level differences in gender composition are mostly driven by shocks to gender composition or skills at the birth cohort level. In this case, our model implies that even without any gender peer effects, the share of girls correlates positively with the academic skills of girls and negatively with the academic skills of boys at the grade level. As a result, the grade level approach overestimates the gender peer effect of girls on girls' academic achievement and underestimates the gender peer effect of girls on boys' academic achievement.

Table 5 replicates table 2 at the grade level to see whether there are signs of non-balancedness in this case. The left-hand-side variable of the regressions underlying table 5 is the individual predicted test score also used in table 2. The right-hand-side variables are school fixed effects, school year fixed effects, and the share of girls at the grade level. For the two larger samples in panels A and C, there is now evidence of statistically significant non-balancedness. The grade level correlation between the predicted test score of girls and the share of girls is significantly positive in the sample of schools with at most two classes per grade and the grade level correlation between the predicted test score of boys and the share of girls is significantly negative in the sample of all schools.[37]

[36]We do not include the school year starting in 2010 to keep the empirical analysis using the grade level approach as close as possible to that using the birth cohort approach. When we include the 2010 school year, the results are very similar to those in table 4 except that estimates become more precise and that the (positive) effect of girls on girls' academic achievement in schools with one class per grade is often statistically significant. It is interesting to note that the effect of the share of girls in a grade on the test performance of girls in table 4 is statistically significant in the sample with all schools, while the birth cohort approach yielded statistically insignificant effects for all schools. As already mentioned, if gender peer effects are at the classroom level, the results using the birth cohort approach can be explained by the share of girls at the birth cohort level being a noisy measure of the share of girls at the classroom level in schools with several classes per grade. On the other hand, the grade level approach will produce the spurious gender peer effects described in section 2.2 no matter how many classes there are per grade as long as grade level differences in gender composition are mostly driven by shocks to gender composition or skills at the birth cohort level.

[37]The signs of the effects in table 5 are consistent with our theoretical model's explanation for the difference between grade level and birth cohort estimates of gender peer effects. When assessing the magnitude it should be kept in mind that much of students' test performance cannot be predicted by the individual characteristics we observe (they account for around 11 to 15 percent of the variance in test performance). When we include the 2010 school year, the results in table 5 change in that a greater share of girls correlates significantly negatively with the predicted test scores of boys in all samples.

5 Counterfactual Analysis

As rates of grade retention in Spain are high compared to the OECD and EU average, it is natural to wonder whether the birth cohort approach could yield a different pattern of gender peer effects than the grade level approach when retention rates are around the OECD and EU average. We examine this question using counterfactual experiments based on a calibrated model. The basis is our theoretical model of schools with grade retention. We first calibrate the model to match our gender-peer-effect estimates for schools with at most two classes per grade in tables 3 and 4. Then we lower the academic thresholds for grade retention in the model – which implies that rates of grade retention fall – and simulate data. This allows us to estimate gender peer effects with the birth cohort and the grade level approach at successively lower rates of grade retention.

5.1 From the Theoretical to the Simulated Model

Our simulations are based on the model described in section 2 and summarized in figure 1. All students in birth cohort t start school in year t and spend the first L years of primary school together in lower grades (LG). After L years in LG, students in birth cohort t have academic skills a^t_{igs} drawn from a school, gender, and birth cohort specific distribution F^t_{gs}. Students with skills a^t_{igs} above the academic threshold p^t_{gs} are promoted to high grade (HG). Students with skills below the academic threshold spend one extra year in LG before being admitted to HG.[38] After one year in HG, students leave primary school.

In the model in section 2 we took the distribution F^t_{gs} to be uniform. Now we can be more general and allow for three alternative skill distributions. On the other hand, we need to assume that the skills students accumulate in HG only depend on the school and on gender, $\omega^\tau_{igs} = \omega_{gs}$.[39] This implies that the HG test scores of students in birth cohort t who have not been retained during primary school are $a^t_{igs} + \omega_{gs}$ with a^t_{igs} drawn from a distribution F^t_{gs} truncated below at p^t_{gs} and that the HG test scores of students in birth cohort t who have been retained during primary school are $a^t_{igs} + \omega_{gs} + \delta^t_{gs}$ with a^t_{igs} drawn from F^t_{gs} truncated above at p^t_{gs}. This is equivalent to assuming that the HG test scores of non-retained students in birth cohort t are equal to a^t_{igs} drawn from F^t_{gs} shifted by ω_{gs} and

[38] Hence, our simulated model continues to assume that all students who are retained are retained at the same stage in primary school. According to the information available from the Spanish national statistical institute's (2013) INEbase database, roughly the same share of students is retained at different stages in primary school in the community of Madrid. But as our data do not specify in which grade students were retained, we cannot calibrate a model where students may be retained at any stage in primary school.

[39] We need to make this assumption to be able to calibrate the model as our data do not allow us to calibrate the distribution of academic skills at different stages of primary school.

truncated below at $p_{gs}^t + \omega_{gs}$ and that the HG test scores of retained students are equal to $a_{igs}^t + \delta_{gs}^t$ with a_{igs}^t drawn from F_{gs}^t shifted by ω_{gs} and truncated above at $p_{gs}^t + \omega_{gs}$. We can therefore simplify the notation by absorbing HG skill accumulation into the mean of F_{gs}^t and the academic thresholds for grade retention p_{gs}^t and set $\omega_{gs} = 0$. Gender peer effects in the model (if any) are assumed to affect the academic skills after L years in LG or the thresholds used for grade retention as described in (25) and (26) respectively.

In the simulations there will be 1000 children per school and birth cohort with the number of girls determined by a binomial distribution with a probability ϕ_s^t that a student is female. The number of schools is set equal to the number of schools with at most two classes per grade in our data.

5.2 Baseline Calibration

We need to calibrate (a) the parameters of the skill distributions F_{gs}^t; (b) the academic thresholds for grade retention p_{gs}^t; (c) the changes δ_{gs}^t in skills that come with grade retention; and (d) the probability that a student is female ϕ_s^t. We choose these parameters so that when the calibrated model is used to simulate data and the simulated data is then used to estimate gender peer effects using the grade level and birth cohort approach, the results match what we obtained for schools with at most two classes per grade in tables 3 and 4. To do so, it is useful to note that the key estimation inputs of the grade level and birth cohort approach are (i) the school and gender specific average test scores of girls and boys at the birth cohort and the grade level, which are the left-hand-side variables of the gender-peer-effect regressions; and (ii) the school specific shares of girls at the birth cohort and the grade level, which are the right-hand-side variables. If the data simulated with the calibrated model reproduces these moments, it will reproduce our gender-peer-effect estimates at the birth cohort and the grade level. We therefore calibrate the model parameters to match these moments.

To get there we have to take different approaches depending on the birth cohort. Our data allow us to observe students in the 2003 and 2004 birth cohorts (the two birth cohorts we used to estimate gender peer effects at the birth cohort level) whether they were retained or not. We can therefore calibrate the parameters corresponding to these birth cohorts so as to match the average test score of non-retained girls and boys in each school; the average test score of retained girls and boys in each school; the share of retained girls and boys in each school; and the share of girls in each school.

Our estimates of gender peer effects at the grade level are based on sixth graders in the school years starting in 2008 and 2009. Sixth graders in 2008 who have been retained once come from the 2002 birth cohort. We therefore also have to calibrate the parameters of the

2002 birth cohort (in addition to the parameters of the 2003 and 2004 birth cohorts). But as we lack data on non-retained students from the 2002 birth cohort we need to take a different approach. We set some parameters equal to the average of the calibrated parameters of the 2003 and 2004 birth cohorts and other parameters to match grade level statistics.

As already mentioned, our simulated model employs both the uniform distribution we assumed for the distribution of academic skills in section 2 and three alternative skill distributions. As a result, we end up with four calibrated models that we can use for counterfactual experiments. We now discuss the calibration of each model in more detail.

5.2.1 Uniform Skill Distribution

In the model in section 2 the distribution of skills was uniform $a_{igs}^t \sim U(\alpha_{gs}^t - \theta, \alpha_{gs}^t + \theta)$. The distribution parameters we need to calibrate are the school, birth cohort, and gender specific means α_{gs}^t and the parameter θ governing skill dispersion.

Calibrating the Parameters of the 2003 and 2004 Birth Cohorts We calibrate the school, birth cohort, and gender specific means α_{gs}^t and the parameter θ governing the dispersion of skills jointly with the academic thresholds for grade retention p_{gs}^t to match (i) the share of girls and boys among students of the same gender who were not retained in any grade at the school and birth cohort level, $\overline{\lambda}_{gs}^t$; (ii) the average test score of non-retained students at the school, birth cohort, and gender level, $\overline{E}_{gs}^t(test\,|non\text{-}retained)$; and (iii) the variance of the test score of non-retained students, $\overline{Var}_{gs}^t(test\,|non\text{-}retained)$, averaged across schools, birth cohorts, and gender. Denoting the cumulative distribution of the uniform $U(\alpha_{gs}^t - \theta, \alpha_{gs}^t + \theta)$ by F_{gs}^t, this yields the first three calibration equations

$$1 - F_{gs}^t(p_{gs}^t) = \overline{\lambda}_{gs}^t \tag{34}$$

$$E_{gs}^t(a\,|a \geq p_{gs}^t) = \overline{E}_{gs}^t(test\,|non\text{-}retained) \tag{35}$$

$$\frac{1}{4S}\sum\nolimits_{t,s,g} Var_{gs}^t(a\,|a \geq p_{gs}^t) = \frac{1}{4S}\sum\nolimits_{t,s,g} \overline{Var}_{gs}^t(test\,|non\text{-}retained) \tag{36}$$

where S is the number of schools. The changes δ_{gs}^t in skills that come with grade repetition are calibrated to get the average test score of retained students in the model to match the average test score of retained students (if any) at the school, birth cohort, and gender level in the data, $\overline{E}_{gs}^t(test\,|retained)$,

$$\delta_{gs}^t + E_{gs}^t(a\,|a < p_{gs}^t) = \overline{E}_{gs}^t(test\,|retained)\,.^{40} \tag{37}$$

[40] Our calibration yields positive average and median changes δ_{gs}^tacross schools for girls as well as boys. There are some empirical estimates of the effect of grade retention on academic skills. The findings of Jacob

This yields a system of $12S + 1$ calibration equations with $12S + 1$ parameters, which can be solved in closed form.[41]

The probability ϕ_s^t of a student in school s and birth cohort t being female is set equal to the share of girls in the school and birth cohort.

Calibrating the Parameters of the 2002 Birth Cohort Students who start sixth grade in 2008 and have been retained once come from the 2002 birth cohort. To do counterfactual experiments using the grade level approach to gender peer effects, we therefore need to calibrate the parameters of the 2002 birth cohort. We do not observe the full 2002 birth cohort however, as non-retained students from this birth cohort attended sixth grade outside of the period covered by our data. As a result, we have to take a different approach to that used for the 2003 and 2004 birth cohorts. We proceed in the following way. The dispersion parameter θ for the 2002 birth cohort is set equal to the calibrated value for the 2003 and 2004 birth cohorts. The probability of a student being female is set equal to the average for the same school that we calibrated for the 2003 and 2004 birth cohorts. The parameters of the skill distribution for girls and boys α_{fs}^{2002} and α_{ms}^{2002} are set equal to the averages that we calibrated for the same school and gender for the 2003 and 2004 birth cohorts. The skill changes associated with grade repetition for girls and boys δ_{fs}^{2002} and δ_{ms}^{2002} and the academic thresholds used for grade retention p_{fs}^{2002} and p_{ms}^{2002} are calibrated to match the average test score of retained girls and boys from the 2002 birth cohort as well as the share of (retained) girls and boys from the 2002 birth cohort among sixth graders in the school year starting in 2008 at the school level.

5.2.2 Normal Skill Distribution

We also calibrate a model where the distribution of skills is taken to be normal with mean α_{gs}^t and standard deviation θ, $a_{igs}^t \sim N(\alpha_{gs}^t, \theta^2)$. The calibration follows the same steps as the calibration of the uniform distribution. As there is no closed-form solution for the calibration equations in (34)-(36), we solve the equations numerically.[42]

and Lefgren (2004) for the US suggest a positive short-term effect on skills of grade retention in third grade but no effect of grade retention in sixth grade. Jacob and Lefgren (2009) find that retained sixth graders in the US are less likely to repeat eighth grade than students who were narrowly not retained in sixth grade. Both Jacob and Lefgren's (2009) study and Manacorda's (2012) analysis for Uruguay find that retained students are more likely to drop out in secondary school than narrowly not retained students.

[41] Using the formula for the variance of uniform distributions and (2), (34), and (36) yields that θ is the positive root of $3\sum_{t,s,g} \overline{Var}_{gs}^t \Big/ \sum_{t,s,g} (\overline{\lambda}_{gs}^t)^2$. Given θ, (2) and (4) imply that (34) and (35) are linear in α_{gs}^t and p_{gs}^t.

[42] While these and subsequent numerical solutions could in principle yield multiple solutions, the solutions found appeared to be unique in practice.

5.2.3 A First Skill Distribution Calibrated to Ensure Raw Test Scores in the $[0, 10]$ Range

So far we have calibrated the skill distributions using standard test scores and ignored that standard test scores are obtained from raw test scores that lie between 0 and 10. We now present a first approach that accounts for these bounds on raw test scores.[43]

To do so we assume that students' raw skills z^t_{igs} are generated according to

$$z^t_{igs} = 10\left(\frac{1}{1+\exp(-v^t_{igs})}\right) \text{ with } v^t_{igs} \sim N(\alpha^t_{gs}, \theta^2) \tag{38}$$

which ensures $0 \leq z^t_{igs} \leq 10$. We use (38) to obtain the density and cumulative distribution of students' standard test scores as $a^t_{igs} = (z^t_{igs} - \overline{\mu}^t)/\overline{\sigma}^t$ where $\overline{\mu}^t$ and $\overline{\sigma}^t$ are the mean and standard deviation of raw scores in the year non-retained students from birth cohort t took the test. The parameters α^t_{gs} in (38) as well as p^t_{gs} and θ are then calibrated by solving (34)-(36) for the 2003 and 2004 birth cohorts. As there is no closed-form solution for the calibration equations, we solve the equations numerically. Once we have calibrated α^t_{gs}, p^t_{gs}, and θ for the 2003 and 2004 birth cohorts, we obtain the values of δ^t_{gs} for these birth cohorts using (37). For the parameters of the 2002 birth cohort we proceed as in the case of the uniform distribution.

5.2.4 A Second Skill Distribution Calibrated to Ensure Raw Test Scores in the $[0, 10]$ Range

A second skill distribution that we calibrate to account for the bounds on raw test scores is the triangular distribution $z^t_{igs} \sim T(l^t_{gs}, u^t_{gs}, m^t_{gs})$ where l and u are the lower and upper bound and $m \in [l, h]$ the mode. Raw test scores can be ensured to lie between 0 and 10 by assuming $l \geq 0$ and $u \leq 10$. Standard test scores are then again obtained as $a^t_{igs} = (z^t_{igs} - \widehat{\mu}^t)/\widehat{\sigma}^t$ where $\widehat{\mu}^t$ and $\widehat{\sigma}^t$ are the mean and standard deviation of raw scores in the year non-retained students from birth cohort t took the test. The triangular distribution generally has three birth cohort, school, and gender specific parameters to be calibrated. The calibration has to therefore proceed somewhat differently than in previous cases.

Our calibration of the triangular skill distributions and the academic thresholds for grade retention for the 2003 and 2004 birth cohorts matches the share of non-retained students and the standard test scores of non-retained students by solving (34)-(35). As there is no closed-form solution, we again rely on numerical solutions. We first try to solve these equations for a given birth cohort, school, and gender assuming full support on $[0, 10]$, that is $l = 0$ and

[43] This allows us to check that the results of our counterfactual experiments are not driven by raw test scores outside of the 0 to 10 range.

$u = 10$. In this case the calibration equations (34)-(35) have only two unknown parameters for a given birth cohort, school, and gender (the mode of the triangular distribution m^t_{gs} and the academic threshold for grade retention p^t_{gs}). For birth cohort, school, and gender combinations where (34)-(35) have no solution with full support on $[0, 10]$ we either (i) relax the assumption that the lower bound of the triangular distribution is 0 and solve for the mode as well as the lower bound of the distribution or (ii) we relax the assumption that the upper bound is 10 and solve for the mode as well as the upper bound of the distribution.[44] Once we have obtained a solution for the parameters of the triangular skill distribution and the academic thresholds, we calibrate the values of δ^t_{gs} for the 2003 and 2004 birth cohorts using (37). For the parameters of the 2002 birth cohort we proceed as in the case of the uniform distribution.

5.3 Counterfactual Simulations of Lower Grade Retention Thresholds

We now use the calibrated models to simulate data and estimate gender peer effects with the birth cohort and the grade level approach. All simulations are based on 1000 children per school and birth cohort. The gender of students is determined by a binomial distribution with a school and birth cohort specific probability of a student being a girl ϕ^t_s. Once the gender of students has been determined, we draw their skills a^t_{igs} from the school, birth cohort, and gender specific skill distributions. Students with skills a^t_{igs} above the threshold for grade retention p^t_{gs} are assumed to enter sixth grade in year $t+5$ and achieve a test score a^t_{igs} in the standardized test. Students with skills a^t_{igs} below the threshold for grade retention p^t_{gs} are assumed to enter sixth grade in year $t+6$ and achieve a test score $a^t_{igs} + \delta^t_{gs}$. The student data simulated with the model are then used to estimate gender peer effects with the grade level approach and the birth cohort approach.

The grade level approach relates the average test score of girls and boys in sixth grade to the share of girls in sixth grade. The average test score of girls and boys in sixth grade in the school year starting in $\tau = 2008, 2009$ is obtained as

$$test_{gs\tau} = \mu_{gs\tau} E(test^{\tau-5}_{igs} \,|\textit{non-retained}) + (1 - \mu_{gs\tau}) E(test^{\tau-6}_{igs} \,|\textit{retained}) \tag{39}$$

[44] It is straightforward to determine which of the two assumptions has to be relaxed to find a solution. For each birth cohort, school, and gender we first obtain all pairs (m^t_{gs}, p^t_{gs}) that solve (34) assuming $l = 0$ and $u = 10$. Then we obtain the highest and the lowest $E^t_{gs}(a \,|a \geq p^t_{gs})$ among these combinations. If the highest $E^t_{gs}(a \,|a \geq p^t_{gs})$ is lower than $\overline{E}^t_{gs}(test \,|\textit{non-repeater})$, the average test score of non-retained students in the data is too high to be matched with a triangular distribution with full support and we relax the assumption that the lower bound is zero. If the lowest $E^t_{gs}(a \,|a \geq p^t_{gs})$ is higher than $\overline{E}^t_{gs}(test \,|\textit{non-repeater})$, the average test score of non-retained students in the data is too low to be matched with a triangular distribution with full support and we relax the assumption that the upper bound is ten.

where $\mu_{gs\tau}$ is the simulated share of non-retained girls or boys among sixth graders of the same gender, while $E(test_{igs}^{\tau-5} \,|non\text{-}retained)$ and $E(test_{igs}^{\tau-6} \,|retained)$ are the simulated average test scores of non-retained students from birth cohort $\tau-5$ and retained students from birth cohort $\tau-6$ respectively. The share of girls at the grade level, $girlsh_{s\tau}$, in the simulations is obtained as the number of non-retained girls in birth cohort $\tau-5$ plus the number of retained girls in birth cohort $\tau-6$ divided by the number of sixth graders in the school year starting in τ. Gender peer effects at the grade level are then estimated by regressing within-school changes in the average test score of girls or boys, $test_{gs2009} - test_{gs2008}$, on the corresponding changes in the share of girls in sixth grade, $girlsh_{s2009} - girlsh_{s2008}$.

The birth cohort approach relates the average test score of girls and boys in a birth cohort to the share of girls in the birth cohort. The average test score of girls and boys in birth cohorts $t = 2003, 2004$ is obtained as

$$test_{gs}^{t} = \lambda_{gs}^{t} E(test_{igs}^{t} \,|non\text{-}retained) + (1-\lambda_{gs}^{t}) E(test_{igs}^{t} \,|retained) \tag{40}$$

where λ_{gs}^{t} is the simulated share of non-retained girls or boys among students of the same gender in birth cohort t, while $E(test_{igs}^{t} \,|non\text{-}retained)$ and $E(test_{igs}^{t} \,|retained)$ are the simulated average test scores of non-retained and retained students from birth cohort t respectively. Gender peer effects at the birth cohort level are then estimated by regressing within-school changes in the average test score of girls or boys, $test_{gs}^{2004} - test_{gs}^{2003}$, on the corresponding changes in the share of girls in these birth cohorts.

5.3.1 Baseline Simulation and Estimation

In the simulation baseline, we simulate data assuming that all model parameters are at their calibrated values. These data are then used to estimate gender peer effects with the birth cohort and the grade level approach. The results are in the first rows of tables 6 and 7. We report point estimates and standard errors averaged across 100 simulations.[45] The point estimates are very similar to the empirical results we obtained for schools with at most two classes per grade in tables 3 and 4. This is not surprising of course. As our calibration matches the average test score at the birth cohort, school, and gender level for the 2003 and 2004 birth cohorts, as well as the share of girls in these birth cohorts, our gender-peer-effect estimates at the birth cohort level with the simulated data have to be very close to our empirical results. As our calibration also matches the average test score of sixth grade girls and boys as well as the share of girls in sixth grade in the school years starting in 2008 and 2009, the same has to be true for gender-peer-effect estimates at the grade level. The main

[45] As we have 1000 students per school and birth cohort, results vary little across simulations.

reason why the gender-peer-effect estimates in the simulation baseline do not coincide exactly with our empirical results is that all birth cohorts are of the same size in our simulations, while the size of birth cohorts in our data differs somewhat across schools and years.

5.3.2 Counterfactual Experiments

In our counterfactual experiments we successively lower the academic thresholds used for grade retention by the same amount for girls and boys in all schools and birth cohorts, starting from the calibrated thresholds. The amount by which we lower these thresholds is chosen to hit certain targets for the share of retained boys averaged across all schools (10, 7.5, 5, 2.5, and 0 percent).[46] As we change the academic thresholds used for grade retention by the same amount independently of the share of girls in the birth cohort, these changes can be thought of as affecting the intercept in (26) but not the marginal effect of the share of girls on the threshold (if any). All model parameters except the thresholds for grade retention are held at their baseline values.

The results of our counterfactual experiments for the four different distributions for students' academic skills are summarized in tables 6 and 7.[47] The results using the birth cohort approach to gender peer effects indicate a statistically significant gender peer effect of girls on boys' academic achievement and a statistically insignificant effect of girls on girls' academic achievement in the simulation baseline and all counterfactual experiments. The size of these effects varies little across counterfactual experiments.[48]

In the simulation baseline, the grade level approach yields a different pattern of gender peer effects than the birth cohort approach. The estimates indicate a statistically significant gender peer effect of girls on girls' academic achievement and a statistically insignificant

[46] In the case of the normal distribution we can never achieve a retention rate that is exactly zero and we therefore targeted a retention rate of 0.001 percent.

[47] As the models used for the counterfactual experiments are calibrated to match our estimates of gender peer effects in schools with at most two classes per grade, the retention rate in the simulation baseline differs from the overall retention rate in our data.

[48] To see why gender-peer-effect estimates using the birth cohort approach may vary little across counterfactual experiments it is useful to return to the expression for the birth cohort gender peer effect in (30). As the model underlying our simulations assumes $\pi_g^\omega = 0$, all the variation in the birth cohort gender peer effect across counterfactual experiments comes from the product between the change in skills associated with grade repetition and the marginal effect of the share of girls in a birth cohort on the retention rate. As the marginal effect of the share of girls in a birth cohort on retention rates turns out to vary little across counterfactual experiments, estimates of gender peer effects at the birth cohort level are similar at different retention rates. With $\pi_g^\omega > 0$, changes in the academic thresholds for grade retention would also affect the strength of gender peer effects through the second term on the right-hand side of (30). This term captures that changes in the retention policy/rate affect the strength of gender peer effects using the birth cohort approach because they affect the expected share of girls the average student in a birth cohort is exposed to in higher grades (when some students will have been retained). We need to assume $\pi_g^\omega = 0$ as our data do not allow us to calibrate the distribution of academic skills at different stages of primary school.

effect of girls on boys' academic achievement. This is not surprising as the model calibration targets the gender-peer-effect estimates for schools with at most two classes per grade in tables 3 and 4. The main finding in tables 6 and 7 is that the grade level approach continues to yield a statistically significant gender peer effect of girls on girls' academic achievement in the counterfactual experiments where the average retention rate for girls and boys is around or somewhat below the average retention rate in the EU or OECD (7 and 8 percent respectively). The grade level approach also continues to indicate a statistically insignificant effect of girls on boys' academic achievement at average EU or OECD retention rates for all distributions except the uniform. Hence, the results in tables 6 and 7 suggest that the birth cohort and the grade level approach may yield different patterns of gender peer effects at rates of grade retention around or below the OECD and EU average. Grade level and birth cohort estimates of gender peer effects coincide when the academic thresholds for grade retention are so low that retention rates are zero (the two approaches become identical in this case).

6 Conclusions

Schools are the obvious place to look for gender peer effects that may arise when girls and boys learn together. The selection of students with different skills into different schools can be bypassed by exploiting within-school variation in gender composition as in Hoxby (2000) and Lavy and Schlosser (2011). Our main theoretical objective has been to understand within-school selection issues that may arise in school systems allowing for grade retention. We found that grade retention generally leads to spurious gender peer effects in academic achievement when gender peer effects are estimated by exploiting within-school differences in gender composition at a given grade level in different years. The direction of spurious gender peer effects depended mainly on the source of shocks to grade level gender composition and the impact of grade repetition on student skills.

Because of the limitations of the grade level approach in school systems with grade retention, we have proposed estimating gender peer effects in academic achievement by exploiting within-school differences in gender composition across birth cohorts. Students are assigned to the same birth cohort if they should have started school in the same year according to the school system's enrollment rule. The birth cohort approach examines whether girls or boys in a birth cohort with a greater share of girls do better academically than students of the same gender in other birth cohorts in the same school. A key feature of the approach is that it does not yield spurious gender peer effects when there is grade retention.

Our main empirical objective has been to estimate gender peer effects in Spanish primary schools, where rates of grade retention are above the OECD and EU average. The birth cohort approach yielded a statistically significant positive gender peer effect of girls on boys' academic achievement by the end of primary school but a statistically insignificant effect of girls on girls' achievement (or equivalently, a statistically significant negative gender peer effect of boys on boys' academic achievement but a statistically insignificant effect of boys on girls' achievement). The grade level approach to gender peer effects tended to yield a different pattern of gender peer effects. The difference between the birth cohort and grade level estimates of gender peer effects turned out to be consistent with our theoretical model if grade repetition improves retained students' skills but students retained in the past still tend to do worse on average than non-retained students in the same grade and if grade level differences in gender composition are mostly driven by shocks to gender composition or skills at the birth cohort level.

Appendix

To prove the results around the inequalities in (11), (18), and (24), we first linearize the within-school grade level change in the share of girls in HG, $\Delta girlsh_{s\tau} = girlsh_{s\tau} - girlsh_{s,\tau-1}$, around $\phi_s^t = 1/2$, $\alpha_{gs}^t = \alpha$, and $p_{gs}^t = p$. Making use of (2) and (9), this yields

$$\text{(A1)} \qquad \begin{aligned} \Delta girlsh_{s\tau} &= \left(\Delta\lambda_{fs}^{\tau-L} - \Delta\lambda_{ms}^{\tau-L} - \Delta\lambda_{fs}^{\tau-L-1} + \Delta\lambda_{ms}^{\tau-L-1}\right)/4 \\ &\quad + \lambda\Delta\phi_s^{\tau-L} + (1-\lambda)\Delta\phi_s^{\tau-L-1} \end{aligned}$$

where $\lambda = (\alpha + \theta - p)/2\theta$, $\Delta\lambda_{gs}^t = (\alpha_{gs}^t - p_{gs}^t - \alpha_{gs}^{t-1} + p_{gs}^{t-1})/2\theta$, and $\Delta\phi_s^t = \phi_s^t - \phi_s^{t-1}$. Linearizing the within-school grade level change in the average test score of girls and boys in HG, $\Delta test_{gs\tau} = test_{gs\tau} - test_{gs,\tau-1}$, using (4)-(9) yields

$$\begin{aligned} \text{(A2)} \qquad \Delta test_{gs\tau} &= \lambda\left(\Delta\alpha_{gs}^{\tau-L} + \Delta p_{gs}^{\tau-L-1}\right)/2 + (1-\lambda)(\Delta\alpha_{gs}^{\tau-L-1} + \Delta p_{gs}^{\tau-L-1})/2 \\ &\quad + (\theta-\delta)\left[2(1-\lambda)\lambda(2\phi_s^{\tau-L-1} - \phi_s^{\tau-L} - \phi_s^{\tau-L-2}) + (1-\lambda)\Delta\lambda_{gs}^{\tau-L} + \lambda\Delta\lambda_{gs}^{\tau-L-1}\right] \end{aligned}$$

where $\Delta\alpha_{gs}^t = \alpha_{gs}^t - \alpha_{gs}^{t-1}$ and $\Delta p_{gs}^t = p_{gs}^t - p_{gs}^{t-1}$.

The standard formula for the least squares regression slope implies that the sign of the least squares slope when regressing (A2) on (A1) across schools is equal to the sign of the covariance between $\Delta girlsh_{s\tau}$ and $\Delta test_{gs\tau}$ across schools. We therefore proceed to calculate this covariance under the different assumptions underlying the results around inequalities in (11), (18), and (24). For example, the inequality in (11) was derived assuming $\alpha_{gs}^t = \alpha$, $p_{gs}^t = p$ and hence $\lambda_{gs}^t = \lambda$. In this case (A1) simplifies to $\Delta girlsh_{gs\tau} = \lambda\phi_s^{\tau-L} - \lambda\phi_s^{\tau-L-1} + (1-\lambda)\phi_s^{\tau-L-1} - (1-\lambda)\phi_s^{\tau-L-2}$ and (A2) to $\Delta test_{gs\tau} = 2(\theta-\delta)(1-\lambda)\lambda(2\phi_s^{\tau-L-1} - \phi_s^{\tau-L} - \phi_s^{\tau-L-2})$. Hence, the assumption of i.i.d. shocks to the share of girls in a birth cohort implies $Cov(\Delta girlsh_{s\tau}, \Delta test_{fs\tau}|\,\tau) = 6(\theta-\delta)(2\lambda-1)(1-\lambda)\lambda Var(\eta)$. As (3) implies $0 < \lambda < 1$, it follows that a least squares regression of within-school changes of HG girls' average test scores on within-school changes of the share of girls in HG yields a strictly positive least squares slope if $(\theta-\delta)(2\lambda-1)Var(\eta) > 0$, which proves (11).

The inequality in (18) was derived assuming $p_{gs}^t = p$ and $\phi_s^t = 1/2$. Substituting in (A1) and (A2) yields $Cov(\Delta girlsh_{s\tau}, \Delta test_{gs\tau}|\,\tau) = 3(2\lambda-1)\delta(1-\rho_\varepsilon)Var(\varepsilon)/16\theta^2$. As $\theta > 0$, this implies a strictly positive least squares slope when regressing (A2) on (A1) across schools if $(2\lambda-1)\delta(1-\rho_\varepsilon)Var(\varepsilon) > 0$, which proves (18). The result in (24) can be proven analogously.

References

Bedard, K. & Dhuey, E. (2006). The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects. *Quarterly Journal of Economics*, *121(4)*, 1437–1472.

Caldera Sanchez, A. & Andrews, D. (2011). To Move or Not to Move: What Drives Residential Mobility Rates in the OECD. Working Paper 846, OECD.

Calsamiglia, C. & Guell, M. (2013). The Illusion of Choice: Evidence from Barcelona. Mimeo. Autonomous University of Barcelona.

European Commission (2011). *Grade retention during compulsory education in Europe: Regulations and statistics*. Education, Audiovisual and Culture Agency, European Commission.

Eurostat (2013). *Regional Yearbook*. Eurostat, European Commission.

Graham, B. S. (2011). Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers. In Jess Benhabib, A. B. & Jackson, M. O. (Eds.), *Handbook of Social Economics*, Volume 1B (pp. 965 – 1052). North-Holland.

Graham, B. S., Imbens, G. W. & Ridder, G. (2010). Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach. Working Paper 16499, National Bureau of Economic Research.

Graham, B. S., Imbens, G. W. & Ridder, G. (2014). Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis. *Quantitative Economics*, *5(1)*, 29–66.

Hanushek, E., Kain, J. & Rivkin, S. (2004). Disruption Versus Tiebout Improvement: The Costs and Benefits of Switching Schools. *Journal of Public Economics*, *88(9-10)*, 1721–1746.

Hoxby, C. (2000). Peer Effects in the Classroom: Learning from Gender and Race Variation. Working paper 7867, National Bureau of Economic Research.

Jacob, B. & Lefgren, L. (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *The Review of Economics and Statistics*, *86(1)*, 226–244.

Jacob, B. & Lefgren, L. (2009). The Effect of Grade Retention on High School Completion. *American Economic Journal: Applied Economics*, *1(3)*, 33–58.

Lavy, V. & Schlosser, A. (2011). Mechanisms and Impacts of Gender Peer Effects at School.

American Economic Journal: Applied Economics, *3(2)*, 1–33.

Manacorda, M. (2012). The Cost of Grade Retention. *The Review of Economics and Statistics*, *94(2)*, 596–606.

PISA (2009a). *The PISA International Database*. OECD.

PISA (2009b). *School Sampling Preparation Manual*. OECD.

Sacerdote, B. (2011). Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far? In Hanushek, E., Machin, S. & Woessmann, L. (Eds.), *Handbook of the Economics of Education*, Volume 3 (pp. 249–277). Elsevier.

Spanish National Statistical Institute (2013). INEbase.

Texas Education Agency (1999, 2011). *Grade Level Retention in Texas Public Schools*. Office of Policy Planning and Research.

UNESCO (2002). Education for All: Is the World on Track? UNESCO.

Whitmore, D. (2005). Resource and Peer Impacts on Girls. *American Economic Review, Papers and Proceedings*, *95(2)*, 199–203.

Figure 1: A School System with Grade Retention

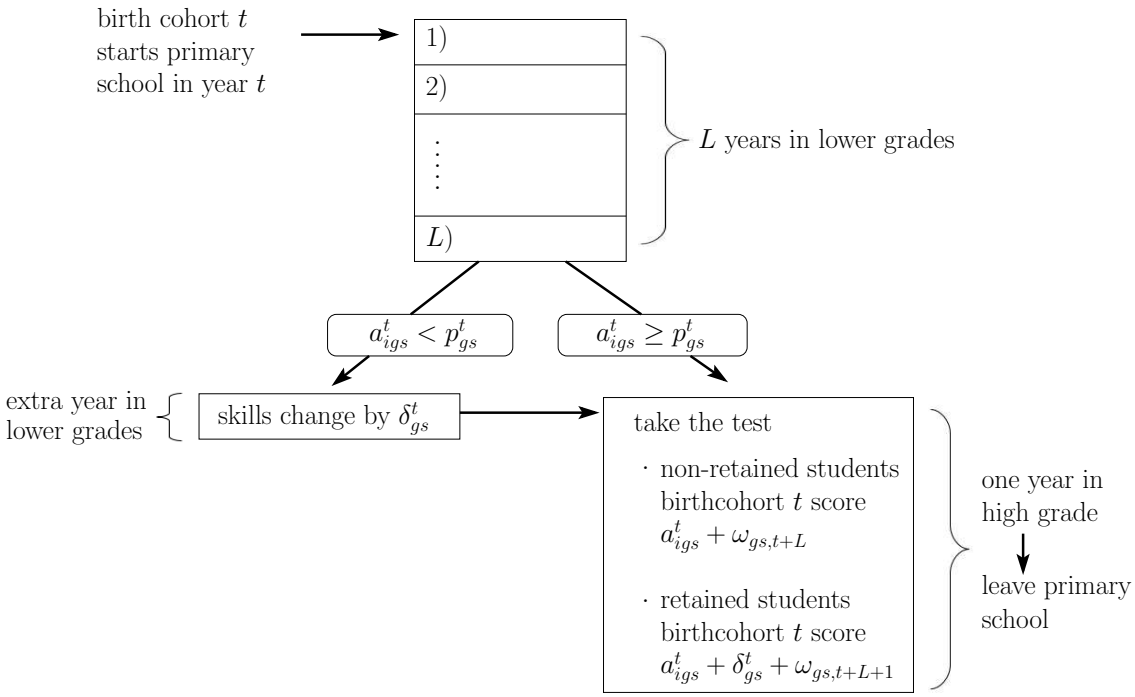

Table 1: Balancedness at the Birth Cohort Level (Schools with at Most Two Classes per Grade)

	Boys	Girls		Boys	Girls
Father profession			Mother profession		
Profession 1	0.003	0.012	Profession 1	0.002	-0.008
	(0.017)	(0.015)		(0.006)	(0.005)
Profession 2	0.011	-0.025	Profession 2	0.058	-0.035
	(0.040)	(0.038)		(0.037)	(0.037)
Profession 3	0.079*	-0.001	Profession 3	-0.019	-0.040
	(0.044)	(0.043)		(0.044)	(0.043)
Profession 4	-0.025	0.002	Profession 4	0.038	-0.011
	(0.012)	(0.012)		(0.030)	(0.029)
Profession 5	-0.059	-0.023	Profession 5	-0.030	0.113**
	(0.043)	(0.046)		(0.043)	(0.046)
Profession 6	-0.065	0.034	Profession 6	0.000	0.017*
	(0.046)	(0.046)		(0.010)	(0.008)
Profession 7	0.018	-0.017	Profession 7	0.004	-0.003
	(0.031)	(0.029)		(0.019)	(0.022)
Profession 8	0.038	0.027	Profession 8	-0.054	-0.033
	(0.029)	(0.029)		(0.053)	(0.052)
Father education			Mother education		
No degree	0.023	-0.015	No degree	0.008	0.014
	(0.032)	(0.036)		(0.031)	(0.038)
Basic secondary	0.041	-0.029	Basic secondary	0.052	-0.090
	(0.058)	(0.059)		(0.056)	(0.057)
Advanced sec.	-0.050	0.013	Advanced sec.	-0.018	0.044
	(0.046)	(0.046)		(0.049)	(0.048)
College	-0.013	0.031	College	-0.041	0.032
	(0.049)	(0.049)		(0.049)	(0.048)
Household composition			School starting age		
Mother	0.037	-0.029	Before 3	-0.023	-0.023
	(0.046)	(0.047)		(0.060)	(0.057)
Father	-0.073	-0.020	Between 3 and 5	0.031	0.030
	(0.082)	(0.097)		(0.056)	(0.056)
One sibling	0.130	-0.173	At 6	0.011	-0.007
	(0.110)	(0.118)		(0.019)	(0.018)

(Continued on next page)

(Continued from previous page)

	Boys	Girls		Boys	Girls
More siblings	-0.126	0.077	7 or more	-0.008	0.000
	(0.084)	(0.086)		(0.013)	(0.011)
Other family	-0.023	0.021	Immigrant	0.040	0.022
	(0.068)	(0.064)		(0.047)	(0.042)
Other	0.044	-0.023	Disability	0.018	-0.032**
	(0.037)	(0.035)		(0.019)	(0.016)
Arrival age	0.180	-0.170	Joint F P-value	0.638	0.694
	(0.329)	(0.298)			

Notes: Estimates of the effect of the share of girls in a birth cohort on listed characteristics, see (31) in the main text for the estimating equation. Robust standard errors clustered at school level in parentheses. The null hypothesis of the joint F-statistic is that the effect of the share of girls in the birth cohort on all listed characteristics is zero. Professions: (1) military, (2) manager or civil servant, (3) professional or technician, (4) secretarial, (5) police, fireman, salesperson, hotel or restaurant staff, (6) construction or maintenance, (7) manufacturing, (8) domestic or cleaning service, private security, concierge. Household composition: with whom the student lives. School starting age: at what age pre-school or school was started. Arrival age: at what age the student arrived to Spain (for students born in Spain the value is 0). Immigrant: student and at least one parent were born outside of Spain. There are 908 schools with at most two classes per grade. *, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.

Table 2: Balancedness of Predicted Test Scores at the Birth Cohort Level

	Boys	Girls
Panel A: Two-Class Schools		
Average score	-0.007	0.007
	(0.034)	(0.038)
Math score	-0.005	0.008
	(0.029)	(0.033)
Panel B: One-Class Schools		
Average score	-0.045	-0.049
	(0.045)	(0.046)
Math score	-0.038	-0.042
	(0.038)	(0.038)
Panel C: All Schools		
Average score	-0.020	-0.053
	(0.054)	(0.060)
Math score	0.002	-0.020
	(0.045)	(0.057)

Notes: Estimates of the effect of the share of girls in a birth cohort on predicted individual test scores. All regressions include school and birth cohort fixed effects. Test scores are predicted separately for girls and boys based on regressions of test scores on all characteristics in table 1 and missing data dummies plus school and birth cohort fixed effects. The estimating equation is (31) in the main text. Robust standard errors clustered at the school level in parentheses. There are 1155 schools in total; 908 schools with at most two classes per grade; and 331 with one class per grade. *, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.

Table 3: Gender Peer Effects Using the Birth Cohort Approach

	Panel A: All Schools					
	Boys	Girls	Boys	Girls	Boys	Girls
Average score	0.150	0.085	0.155	0.085	0.155	0.067
	(0.106)	(0.110)	(0.102)	(0.105)	(0.103)	(0.104)
Math score	0.160	0.090	0.17*	0.090	0.17*	0.076
	(0.104)	(0.111)	(0.102)	(0.109)	(0.103)	(0.108)
Individual controls			yes	yes	yes	yes
Peer group controls					yes	yes
	Panel B: Two-Class Schools					
	Boys	Girls	Boys	Girls	Boys	Girls
Average score	0.238**	0.099	0.255**	0.113	0.253**	0.101
	(0.115)	(0.122)	(0.112)	(0.117)	(0.113)	(0.114)
Math score	0.241**	0.113	0.256**	0.128	0.255**	0.111
	(0.114)	(0.126)	(0.112)	(0.123)	(0.114)	(0.121)
Individual controls			yes	yes	yes	yes
Peer group controls					yes	yes
	Panel C: One-Class Schools					
	Boys	Girls	Boys	Girls	Boys	Girls
Average score	0.263	-0.049	0.313*	0.017	0.317*	0.023
	(0.183)	(0.169)	(0.182)	(0.167)	(0.180)	(0.160)
Math score	0.370**	-0.103	0.397**	-0.037	0.388**	-0.074
	(0.179)	(0.183)	(0.179)	(0.179)	(0.184)	(0.169)
Individual controls			yes	yes	yes	yes
Peer group controls					yes	yes

Notes: Estimates of the effect of the share of girls in a birth cohort on the test scores of boys and girls. The estimating equation is (32) in the main text. Robust standard errors clustered at the school level in parentheses. Individual controls refer to the individual characteristics in table 1 and missing data dummies. Peer group controls are the shares of mothers and of fathers in the birth cohort with the four educations levels in table 1, the share of immigrants in the birth cohort, and the size of the birth cohort. There are 1155 schools in total; 908 schools with at most two classes per grade; and 331 with one class per grade. *, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.

Table 4: Gender Peer Effects Using the Grade Level Approach

	Panel A: All Schools					
	Boys	Girls	Boys	Girls	Boys	Girls
Average score	-0.088	0.315***	-0.061	0.288***	-0.047	0.274**
	(0.121)	(0.121)	(0.118)	(0.118)	(0.118)	(0.118)
Math score	-0.057	0.299**	-0.035	0.274**	-0.023	0.260**
	(0.118)	(0.126)	(0.115)	(0.123)	(0.115)	(0.124)
Individual controls			yes	yes	yes	yes
Peer group controls					yes	yes
	Panel B: Two-Class Schools					
	Boys	Girls	Boys	Girls	Boys	Girls
Average score	0.058	0.406***	0.091	0.381***	0.101	0.355***
	(0.135)	(0.134)	(0.130)	(0.131)	(0.129)	(0.130)
Math score	0.030	0.349**	0.057	0.326**	0.067	0.295**
	(0.130)	(0.142)	(0.126)	(0.140)	(0.126)	(0.140)
Individual controls			yes	yes	yes	yes
Peer group controls					yes	yes
	Panel C: One-Class Schools					
	Boys	Girls	Boys	Girls	Boys	Girls
Average score	-0.042	0.222	0.076	0.267	0.046	0.200
	(0.201)	(0.195)	(0.198)	(0.189)	(0.199)	(0.185)
Math score	-0.081	0.121	0.011	0.160	-0.004	0.088
	(0.202)	(0.214)	(0.198)	(0.207)	(0.199)	(0.203)
Individual controls			yes	yes	yes	yes
Peer group controls					yes	yes

Notes: Estimates of the effect of the share of girls in sixth grade on the test scores of boys and girls. The estimating equation is (33) in the main text. Robust standard errors clustered at the school level in parentheses. Individual controls refer to the individual characteristics in table 1 and missing data dummies. Peer group controls are the shares of mothers and of fathers in the grade with the four educations levels in table 1, the share of immigrants in the grade, and the number of students in the grade. There are 1155 schools in total; 908 schools with at most two classes per grade; and 331 with one class per grade. *, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.

Table 5: Balancedness of Predicted Test Scores at the Grade Level

	Boys	Girls
Panel A: Two-Class Schools		
Average score	-0.051	0.058*
	(0.039)	(0.033)
Math score	-0.042	0.059**
	((0.033)	(0.030)
Panel B: One-Class Schools		
Average score	-0.041	0.033
	(0.050)	(0.047)
Math score	-0.058	-0.006
	(0.037)	(0.037)
Panel C: All Schools		
Average score	-0.098**	0.020
	(0.049)	(0.050)
Math score	-0.087*	0.030
	(0.050)	(0.049)

Notes: Estimates of the effect of the share of girls in sixth grade on predicted individual test scores. All regressions include school and school year fixed effects. Test scores are predicted separately for girls and boys based on regressions of test scores on all characteristics in table 1 and missing data dummies plus school and grade fixed effects. The estimating equation is as in (31) in the main text but at the grade level. Robust standard errors clustered at the school level in parentheses. There are 1155 schools in total; 908 schools with at most two classes per grade; and 331 with one class per grade. *, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.

Table 6: Counterfactual Experiments of Lower Thresholds for Grade Retention Based on the Uniform and Normal Distribution of Skills

Retention Rate		Uniform Distribution				Normal Distribution			
		Grade		Birth Cohort		Grade		Birth Cohort	
Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
0.189	0.156	0.09	0.41	0.23	0.10	0.09	0.42	0.22	0.10
		(0.13)	(0.12)	(0.12)	(0.12)	(0.13)	(0.12)	(0.12)	(0.12)
0.100	0.075	0.20	0.30	0.25	0.08	0.04	0.50	0.24	0.10
		(0.12)	(0.12)	(0.13)	(0.13)	(0.13)	(0.13)	(0.12)	(0.13)
0.075	0.054	0.21	0.31	0.26	0.09	0.06	0.47	0.25	0.10
		(0.13)	(0.12)	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)
0.050	0.034	0.26	0.32	0.27	0.09	0.10	0.41	0.26	0.10
		(0.13)	(0.13)	(0.13)	(0.14)	(0.13)	(0.13)	(0.13)	(0.13)
0.025	0.016	0.29	0.27	0.29	0.10	0.17	0.31	0.27	0.11
		(0.14)	(0.13)	(0.14)	(0.14)	(0.14)	(0.14)	(0.14)	(0.14)
0.000	0.000	0.27	0.13	0.27	0.13	0.27	0.13	0.27	0.13
		(0.15)	(0.15)	(0.15)	(0.15)	(0.14)	(0.14)	(0.14)	(0.14)

Notes: The coefficients and standard errors in parentheses are averages across 100 simulations.

Table 7: Counterfactual Experiments of Lower Thresholds for Grade Retention Based on Distributions Ensuring Raw Skills in the 0-10 Range

Retention Rate		$[0, 10]$ Distribution				Triangular Distribution			
		Grade		Birth Cohort		Grade		Birth Cohort	
Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
0.189	0.157	0.04	0.45	0.22	0.11	0.06	0.45	0.23	0.11
		(0.13)	(0.13)	(0.11)	(0.12)	(0.13)	(0.13)	(0.11)	(0.11)
0.100	0.078	0.00	0.54	0.23	0.11	0.07	0.46	0.26	0.12
		(0.13)	(0.13)	(0.12)	(0.12)	(0.14)	(0.14)	(0.13)	(0.13)
0.075	0.057	0.02	0.52	0.23	0.11	0.10	0.40	0.26	0.12
		(0.13)	(0.13)	(0.12)	(0.12)	(0.14)	(0.14)	(0.12)	(0.13)
0.050	0.036	0.06	0.46	0.23	0.11	0.16	0.36	0.27	0.12
		(0.13)	(0.13)	(0.12)	(0.12)	(0.14)	(0.14)	(0.13)	(0.13)
0.025	0.018	0.13	0.35	0.24	0.11	0.22	0.30	0.26	0.12
		(0.13)	(0.13)	(0.13)	(0.13)	(0.14)	(0.14)	(0.14)	(0.14)
0.000	0.000	0.24	0.12	0.24	0.12	0.27	0.12	0.27	0.12
		(0.13)	(0.13)	(0.13)	(0.13)	(0.14)	(0.13)	(0.14)	(0.13)

Notes: The coefficients and standard errors in parentheses are averages across 100 simulations.

Appendix Table 1: Descriptive Statistics

	2008			2009			2010		
	All	Boys	Girls	All	Boys	Girls	All	Boys	Girls
Summary of raw test scores									
Average score	5.34	5.36	5.33	6.44	6.50	6.39	6.52	6.56	6.48
(Standard deviation)	(2.24)	(2.27)	(2.21)	(2.24)	(2.29)	(2.19)	(2.04)	(2.08)	(2.00)
Math score	5.00	5.17	4.81	5.45	5.65	5.23	6.07	6.16	5.98
(Standard deviation)	(2.53)	(2.54)	(2.51)	(2.74)	(2.73)	(2.73)	(2.80)	(2.79)	(2.80)
Retained once (%)	15	16	13	14	15	13	15	16	13
Retained more than once (%)	0.4	0.5	0.3	0.4	0.4	0.4	0.6	0.7	0.4
Immigrants (%)	19	19	19	18	17	18	17	17	17
Father education (%)									
No degree	10	9	10	9	9	9	8	8	9
Basic secondary	35	35	35	36	36	37	36	36	36
Advanced secondary	17	17	17	17	17	18	18	18	18
College	38	39	38	37	38	36	38	39	37
Mother education (%)									
No degree	9	8	10	8	7	9	7	7	8
Basic secondary	32	32	33	34	33	34	33	32	34
Advanced secondary	18	18	18	19	19	19	18	19	18
College	39	40	37	38	40	37	40	41	39

Notes: Average score is an average of mathematics, reading, general knowledge, and dictation.