Peer Effects when Girls and Boys Learn Together,  
a Birth Cohort Approach

Antonio Ciccone  Walter Garcia-Fontes*

April 2015

Abstract

We propose estimating peer effects when girls and boys learn together by exploiting within-school variation in the share of girls/boys across birth cohorts. In school systems without grade retention, our birth cohort approach is equivalent to the grade level approach used in the literature. In school systems with grade retention, our birth cohort approach has the advantage that it does not yield spurious peer effects of girls/boys on the learning success of girls or boys. We apply the birth cohort approach and the grade level approach to primary schools in Spain, which permit grade retention.

Keywords: Peer effects in school, grade retention, selection bias

JEL Codes: I21, J16, J24

*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 data on student mobility in public schools, and Carmen Arias and Leticia Moreno for their help in collecting additional data on student mobility. We gratefully acknowledge financial support from CREI, fede, 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). The paper was previously entitled "Gender peer effects in school, a birth cohort approach."
1 Introduction

Do girls learn more when they learn together with girls instead of boys? And what about boys? Peer effects that arise when girls and boys learn together have been debated since the introduction of coeducation and matter for the design of school systems as well as the policy response to the recent revival of single-sex schools (Hoxby, 2000; Whitmore, 2005; Lavy and Schlosser, 2011; Lee et al., 2014).\(^1\) 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 (2000), Lavy and Schlosser (2011), and Lee et al. (2014). Lee et al. (2014) exploit random assignment of students to schools and classrooms in Korea.\(^2\) Hoxby (2000) and Lavy and Schlosser (2011) bypass selection into different schools by examining the response of academic achievement to within-school differences in gender composition at a given grade level in Texas and Israel. Examining the consequences of within-school differences in gender composition is appealing as they partly reflect natural variation in the births of girls and boys in school catchment areas.

We develop an alternative approach to estimate peer effects when girls and boys learn together. Our approach exploits within-school variation in the share of girls/boys across birth cohorts. In school systems that do not practice grade retention or delayed enrollment of students with weak cognitive or academic skills, the birth cohort approach is equivalent to the within-school grade level approach of Hoxby (2000) and Lavy and Schlosser (2011). In school systems with grade retention or delayed enrollment, the birth cohort approach has the advantage that it does not yield spurious peer effects of girls/boys on the learning success of girls or boys. In contrast, the grade level approach generally yields spurious gender peer effects in school systems with grade retention, even if grade level differences in gender composition were solely driven by natural variation in the share of girls across birth cohorts. This is because of selection effects which cannot be avoided by controlling for (endogenous) grade retention at the individual level.

Most school systems allow for grade retention of students who have not acquired certain academic skills or achieved learning goals by the end of the academic year. For example, twenty-four of twenty-seven European Union countries permit grade retention starting in primary school (European Commission, 2011). Also, there is grade retention starting in primary school in two of the three countries for which we currently have estimates of gender peer effects (Israel and the United States; there is no grade retention in Korea). 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 (PISA, 2009a).\(^3\)

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

\(^2\) Lee et al. (2014) also exploit random assignment to single-sex schools in Korea. 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
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 United States (10.6 percent). Retention rates in primary school are lower in Canada and Israel (4.2 percent) and the UK (1.7 percent). An advantage of our birth cohort approach to gender peer effects is that it can be applied in these countries as well as countries that do not practice grade retention.

Our birth cohort approach to peer effects when girls and boys learn together is based on within-school differences in the share of girls/boys across birth cohorts. Students in a school 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 then examines whether girls or boys learn more if they belong to a birth cohort with a greater share of girls. As students in the same birth cohort may not be in the same grade because of grade retention or delayed enrollment, gender peer effects estimated using the birth cohort approach are intention-to-treat effects. Two key features of the birth cohort approach to gender peer effects are (i) that it does not yield spurious gender peer effects when there is grade retention or delayed enrollment and (ii) that the approach is equivalent to the grade level approach when grade retention is not permitted and students start school according to the school system’s enrollment rule.

We use the birth cohort approach to estimate peer effects when girls and boys learn together in a sample of primary schools in Spain, which allows for grade retention. An advantage of our study compared to existing studies of gender peer effects in school systems with grade retention – Hoxby’s (2000) study for Texas and Lavy and Schlosser’s (2011) study for Israel for example – is that our data allow estimating gender peer effects at the birth cohort level as well as the grade level. Another advantage is that we have data on almost 40 student characteristics, which enables us to check the balancedness of student characteristics with respect to the share of girls in a birth cohort or grade in detail. We find the characteristics of students, and the academic achievement predicted by student characteristics, to be unrelated to the share of girls at the birth cohort level (at the grade level, we do find predicted achievement and the share of girls to be related).

When we estimate gender peer effects using the birth cohort approach, we find a statistically significant, positive peer effect of girls on the individual 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.\footnote{Girls do worse than boys on average and this effect can therefore not be explained by spillovers from high-skill to low-skill students. 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.} On the other hand, the effect of the share of girls in a birth cohort on the individual academic achievement of girls is statistically insignificant. The grade level approach tends to yield the opposite pattern of gender peer effects. The effect of the share of girls in a grade on the individual academic achievement of girls is positive and statistically significant, while the effect on the individual achievement of boys is statistically insignificant. We show that the difference between the birth cohort and the grade level approach is consistent with a theoretical model of grade retention as long as (i) grade repetition improves retained students’ skills but students retained in the past (and therefore negatively selected) 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.

The share of students subject to grade retention in Spain is above the OECD and EU average. It is therefore 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 of grade retention. To do so, we first calibrate the model to match our estimates for Spain. Then we lower the academic thresholds used for grade retention in the model to achieve lower rates of grade retention. Finally we use the model with lower retention thresholds to simulate data, and estimate gender peer effects with the simulated data. 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 below the OECD and EU average.

Our data specify the primary school that students attended when their skills were tested but not where students went to school previously. This is also true for Hoxby’s (2000) data for Texas, Lavy and Schlosser’s (2011) data for Israel, and Lee et al.’s (2014) data for Korea. Student mobility across schools in Israel and Korea 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 and Lee et al. estimate that 5 percent of Korean students in grades six to nine quit school or transferred to another school. On the other hand, Hanushek, Kain, and Rivkin (2004) report an annual rate of student mobility in Texas of 24 percent. In our data, student mobility is 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 a 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 in detail. 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 contributes to the literature on peer effects, see Sacerdote (2011, 2014) for recent reviews. Within this literature we are most closely related to the work of Hoxby (2000), Lavy and Schlosser (2011), and Lee et al. (2014) who estimate gender peer effects in school in Texas, Israel, and Korea, respectively. Lee et al.’s (2014) study uses that in Korea, the assignment mechanism of students to middle schools does not consider preferences over schools and assigns students randomly across schools within school districts and across classrooms within schools. Moreover, as grade retention is not permitted, this random assignment is not undone later by grade retention of students who have not acquired certain academic skills or achieved learning goals. Lee et al.’s estimates of gender peer effects in coed classrooms point into the same direction as our estimates using the birth cohort approach in two respects. First, they also find statistically significant peer effects of girls on boys’ overall achievement, but not of girls on girls’ achievement. Second, the magnitude of the effect of girls on boys’ overall achievement in Lee et al. is similar to our result for schools with only one class per grade (this is our most directly comparable result).

Lavy and Schlosser (2011) estimate gender peer effects using a within-school grade level approach in Israeli primary, middle, and high schools. In primary schools, 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 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 peer effects of girls on girls’ achievement in mathematics and English but not in science and technology or Hebrew. Peer effects of girls on boys’ achievement are statistically insignificant in all subjects. In high school, Lavy and Schlosser find 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 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.

8Their estimates for single-sex classrooms in coed schools also point in the same direction as our estimates. In addition, Lee et al. examine the effects of instruction in single-sex schools. They find that boys in single-sex schools do better than boys in coed schools (while girls do not do better in single-sex than coed schools). One explanation for this finding is that boys have special needs and teachers in boys only schools can specialize.

9They estimate that a 10-percentage-point increase in the share of girls in a classroom improves boys’ overall academic achievement by 3.6 percent of a standard deviation. Our point estimate is that a 10-percentage-point increase in the share of girls improves boys’ achievement by 3.1 percent of a standard deviation.

10For a comparison of the confidence intervals in Lavy and Schlosser and in Hoxby with our confidence intervals, see Ciccone and Garcia-Fontes (2014).
Hoxby’s (2000) study of gender peer effects using a within-school grade level approach in Texas primary schools finds statistically significant effects of girls on girls’ achievement in mathematics and reading in grades three to six. Peer effects of girls on boys’ achievement are also mostly significant in reading, with the exception of fourth grade. 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.

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.\textsuperscript{11} Her approach exploits differences in gender composition across classes 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.\textsuperscript{12} 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 the other studies).\textsuperscript{13}

\subsection*{1.2 Preview of Main Theoretical Findings}

A main finding from our theoretical analysis of school systems with grade retention is that a within-school grade level approach 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 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, so that 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 as long as the gender of retained students joining the class is independent of the gender composition of the class they join.\textsuperscript{14} Now consider the probability that a girl in the class is a student who joined the class because she was retained. This probability should be lower in classes that started out with an

\textsuperscript{11}The project randomly assigned kindergarten children in participating schools and students entering participating schools to classes (of different types).
\textsuperscript{12}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.
\textsuperscript{13}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). Another interesting paper on gender peer effects in school is Lu and Anderson (2015). They examine gender peer effects within subclassroom groups in Chinese middle schools and find that girls who are randomly assigned to sit near girls achieve higher test scores.
\textsuperscript{14}It is hard to see why a school would decide to retain a student or not depending on the gender composition of the class the retained student would end up joining. But it is possible.
above-average share of girls, as the gender of retained students joining the class should be independent of the gender composition of the class they join. As retained students are selected, the distribution of skills of girls will therefore be correlated with the share of girls in the grade. For example, suppose that classes always start with 30 students in first grade and that half of them are girls on average. Suppose also that typically 20 percent of students are retained because of low skills (the argument goes through even if retention rates differ for girls and boys). Now consider a class starting with 20 girls in first grade in 2005. Suppose that the class starting first grade a year earlier (in 2004) had the average number of girls (15). By the time the 2005 class has reached higher grades, 4 girls (0.2*20) will have been retained and 3 retained girls (0.2*15) from the 2004 class will have joined. Hence, 3 of 19 girls in the class will have been retained in the past because of low skills. This implies that in higher grades, the class ends up with 63 percent girls (19/30) and a probability that a girl in the class has been retained because of low skills of 16 percent (3/19). Compared to the average incoming class in the school – with 50 percent girls and a probability that a girl in the class has been retained because of low skills of 20 percent – incoming classes with a greater share of girls therefore end up having more girls and a lower probability that a girl in the class has been retained because of low skills. This is a source of spurious gender peer effects when using a grade level approach. The symmetric argument for boys and the fact that the share of girls and boys sums to one, implies that the probability that a boy in the class has been retained should be higher in classes that started out with an above-average share of girls (or equivalently, that started out with a below-average share of boys).

The direction of the spurious gender peer effects depends on whether non-retained students do better or worse on average than students who were retained in the past. This in turn depends on the balance of two effects. On the one hand, retained students are negatively selected. On the other hand, retained students have had one more year of schooling than non-retained students in the same grade, which may have helped them to catch up in skills. If (as in our data) the negative selection effect dominates and retained students do worse, a class starting with an above-average share of girls in first grade will tend to have both a greater share of girls and girls with better skills in higher grades (although there are no gender peer effects). This positive association between the share of girls and girls’ skills in higher grades produces a spurious positive effect of the share of girls on girls’ achievement when using a grade level approach to gender peer effects. For boys, the argument is symmetric and implies a spurious positive peer effect of the share of boys in a grade on the academic achievement of boys, or equivalently, a spurious negative peer effect of the share of girls on the achievement of boys.

A grade level approach also yields spurious gender peer effects when there are ex-

---

15 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 an above-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 secondary empirical relevance as rates of grade retention are below 50 percent in almost all countries (PISA, 2009a).
ogenous shocks to the skills of girls and boys in a birth cohort. Such skill shocks do not, as a first analysis might suggest, necessarily translate into a spurious positive 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 class where the share of girls is equal to the average in the school but girls have above-average skills. As this class proceeds from first grade to higher grades, an above-average share of girls will be promoted (not retained). Moreover, the skills of non-retained girls will be above average. Hence, in higher grades, this class will end up with an above-average share of girls and girls with above-average skills. This would tend to translate into a spurious positive effect of the share of girls on the academic achievement of girls when using a grade level approach.

However, there is a countervailing force operating in the class one year behind the class starting with girls with above-average skills. As this class proceeds to higher grades, it receives retained girls from the class starting with girls with above-average skills. As the skills of these retained girls are above average and their number below average, the class one year behind the class starting with girls with above-average skills ends up with girls with above-average skills and a below-average share of girls in higher grades. In our framework, this countervailing force is dominated if and only if grade repetition improves retained students’ skills. In this case, shocks to the skills of girls at the birth cohort level translate into a spurious positive peer effect of the share of girls in a grade on the academic achievement of girls. The argument for boys is symmetric and implies a spurious negative grade level effect of girls on boys’ academic achievement if grade repetition improves retained students’ skills.

An interesting follow-up question is whether spurious gender peer effects when using a grade level approach can be avoided by using the share of girls at the birth cohort level as an instrument for the share of girls at the grade level. For such an 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 that this is not the case 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. We will also show that spurious gender peer effects when using a grade level approach can generally not be avoided by controlling for (endogenous) grade retention at the individual level. However, it is possible to avoid spurious gender peer effects when using the grade level approach by simultaneously controlling for grade retention at the individual level and instrumenting the share of girls at the grade level by the share of girls at the birth cohort level.
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 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 \( L \)th 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 \( a_{igt} \) 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

\[
\begin{align*}
  a_{igt} & \geq p_{gs}^t, \\
\end{align*}
\]

Students with skills below the academic threshold, \( a_{igt} < p_{gs}^t \), 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 \( \phi_s^t \) 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 \( \alpha_{gs}^t \).  

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 \( \tau \) is taken to reflect their academic skills \( a_{igt} + w_{igt} \) where \( w_{igt} \) refers to the skills that student \( i \) of gender \( g \) acquires by attending HG in the school year starting in \( \tau \).

---

\(^{16}\)In our simulated models in section 5 we can consider a wider range of distributions.
We assume that the $w_{gs}$ of different students are obtained as independent draws from a distribution with a constant variance and a mean $\omega_{gs}$ 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 $\alpha_{gs} + w_{gs} + \delta_{gs}$, where $\delta_{gs}$ 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 $\delta_{gs} > 0$, 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 in principle affect the academic skills of HG students and hence their test results in three main ways. First, the share of girls among a student’s peers during the first $L$ years of LG could affect the student’s academic skills $\alpha_{gs}$ at the end of the $L$th school year. Second, the share of girls among a student’s peers could affect the threshold $p_{gs}$ used to determine who is promoted to HG after $L$ years in LG. Third, the academic skills $w_{gs}$ a student accumulates in HG could depend on the share of girls among the student’s HG peers.

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

$$\lambda_{gs}^t = \frac{1}{2^l}(\alpha_{gs}^t + \theta - p_{gs}^t)$$

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

$$-\theta < \alpha_{gs}^t - p_{gs}^t < \theta.$$
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 $\tau$ 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

$$\mu_{f,s,\tau} = \frac{\phi_s^{\tau-L} \lambda_s^{\tau-L}}{\phi_s^{\tau-L} \lambda_{f,s}^{\tau-L} + \phi_s^{\tau-L-1}(1 - \lambda_{f,s}^{\tau-L-1})}$$

where $f$ stands for female, $\lambda_{f,s}^{\tau}$ is the share of non-retained girls among all girls in birth cohort $\tau$, and $\phi_s^{\tau}$ is the share of girls among all students in birth cohort $\tau$. The share of non-retained boys among HG boys can be obtained analogously

$$\mu_{m,s,\tau} = \frac{(1 - \phi_s^{\tau-L}) \lambda_{m,s}^{\tau-L}}{(1 - \phi_m^{\tau-L}) \lambda_{m,s}^{\tau-L} + (1 - \phi_m^{\tau-L-1})(1 - \lambda_{m,s}^{\tau-L-1})}$$

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

$$test_{gs,\tau} = \mu_{gs,\tau} E(test_{s,\tau}^{\tau-L} | non-retained) + (1 - \mu_{gs,\tau}) E(test_{s,\tau}^{\tau-L-1} | retained).$$

We now use the model to examine if and why a grade level approach gives rise to gender peer effects even if there are none and to show that spurious gender peer effects at the grade level cannot be eliminated by controlling for (endogenous) grade retention at the individual level. Then we introduce true gender peer effects into the model to examine the birth cohort approach to gender peer effects.

### 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 in fact there are none? And in what direction might spurious gender peer effects go? To address these questions, we assume throughout this section 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.\(^{17}\)

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 $\tau$ and $\tau - 1$. A grade level approach would assess the strength of gender peer effects by regressing the individual test performance of HG girls or boys on the share of girls in HG and school fixed effects. This yields the same asymptotic result as regressing the change in the average test performance of HG girls or boys between the school years starting in $\tau$ and $\tau - 1$, $test_{gs,\tau} - test_{gs,\tau-1}$, on the corresponding change in the share of HG girls, $girlsh_{s,\tau} -$\(^{17}\)That is, we assume that $\alpha_{gs,\tau}, \beta_{gs,\tau}$, and $\omega_{gs}$ do not depend on the share of girls in LG or HG.
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.

$$
girls_{\tau} = \frac{\phi_{s}^{\tau - L} \alpha_{fs}^{\tau - L} + \phi_{s}^{\tau - L - 1} (1 - \lambda_{fs}^{\tau - L} - 1)}{\phi_{s}^{\tau - L} \lambda_{fs}^{\tau - L} + \phi_{s}^{\tau - L - 1} (1 - \lambda_{ms}^{\tau - L} - 1) + (1 - \phi_{s}^{\tau - L}) \lambda_{ms}^{\tau - L} + (1 - \phi_{s}^{\tau - L - 1}) (1 - \lambda_{ms}^{\tau - L} - 1)}.
$$

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 ($\phi_{s}^{\tau}$), shocks to the academic skills of girls and boys at the birth cohort level ($\alpha_{fs}^{\tau}, \alpha_{ms}^{\tau}$), and shocks to the academic thresholds for grade retention ($p_{fs}^{\tau}, p_{ms}^{\tau}$). 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, $\delta_{gs}^{\tau} = \delta$ 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, $\alpha_{fs}^{\tau} = \alpha_{ms}^{\tau} = \alpha$ and $p_{fs}^{\tau} = p_{ms}^{\tau} = p$. The only shocks are i.i.d. shocks $\eta_{s}^{\tau}$ with mean zero to the share of girls in a birth cohort

$$
\phi_{s}^{\tau} = 1/2 + \eta_{s}^{\tau},
$$

where $Var(\eta_{s}^{\tau}) = Var(\eta) > 0$. What would we obtain when estimating gender peer effects in HG using a grade level approach? Linearizing the share of girls in HG in (9) yields that the grade level approach would yield a spurious positive peer effect of girls on girls’ academic achievement and a spurious negative peer effect of girls on boys’ achievement if

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

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 $\lambda$ is the share of non-retained students evaluated at $\alpha$ and $p$. If $(2 \lambda - 1) (\theta - \delta) < 0$, the grade level approach would yield a spurious negative peer effect of girls on girls’ academic achievement and a spurious positive 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 peer effect of girls on girls’ academic achievement and a spurious negative 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

---

18See the appendix for a proof of this result.
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 \( \eta \) 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

\[
\text{girlsh}_{s,t+L} = \lambda \eta \quad \text{and} \quad \text{girlsh}_{s,t+L+1} = (1 - \lambda) \eta
\]

where \( \hat{x} \) denotes deviations of \( x \) from its value in the absence of shocks. These expressions reflect that a share \( \lambda \) of the \( \eta \) 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

\[
\hat{\mu}_{fs,t+L} = 2\lambda(1 - \lambda) \eta.
\]

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)

\[
\hat{\text{test}}_{gs,t+L} = \hat{\mu}_{fs,t+L} (\theta - \delta) \eta = 2\lambda(1 - \lambda) (\theta - \delta) \eta.
\]

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

\[
\hat{\text{test}}_{gs,t+L+1} = -\hat{\mu}_{fs,t+L+1} (\theta - \delta) \eta = -2\lambda(1 - \lambda) (\theta - \delta) \eta.
\]
Combining (12) and (14)-(15) yields that if \( \theta > \delta \), a positive shock \( \eta \) 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 \( \tau \) come either from birth cohort \( \tau - L \) (non-repeaters) or birth cohort \( \tau - L - 1 \) (repeaters). Let \( \eta' \) be the shock to gender composition in birth cohort \( \tau - L \) and \( \eta'' \) 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 \( \tau \) 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

\[
E(\hat{\text{girlsh}}_{\tau \text{test}f,\tau} \mid \tau) = 2\lambda(1 - \lambda)(\theta - \delta) \text{Var}(\eta)
\]

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 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

\[
\alpha'_{fs} = \alpha + \epsilon'_{fs} \quad \text{and} \quad \alpha'_{ms} = \alpha + \epsilon'_{ms}
\]

where \( \epsilon'_{fs}, \epsilon'_{ms} \) are shocks with mean zero with \( \text{Var}(\epsilon'_{fs}) = \text{Var}(\epsilon'_{ms}) = \text{Var}(\epsilon) > 0 \) and \( \text{Correl}(\epsilon'_{fs}, \epsilon'_{ms}) = \rho_\epsilon \). 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 peer effect of girls on girls’ academic achievement and a spurious negative peer effect of girls on boys’ achievement if

\[(2\lambda - 1)\delta > 0.19\]  

If \((2\lambda - 1)\delta < 0\), the grade level approach would yield a spurious negative peer effect of girls on girls’ academic achievement and a spurious positive 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 peer effect of girls on girls’ academic achievement and a spurious negative 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 \(\varepsilon\) to the average skills of girls in birth cohort \(t\) for the share of HG girls and their academic skills 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 \(t + L + 1\),

\[\hat{girlsh}_{t+L} = \frac{1}{8\theta} \varepsilon \quad \text{and} \quad \hat{girlsh}_{t+L+1} = -\frac{1}{8\theta} \varepsilon\]  

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, respectively,

\[\hat{test}_{t+L} = \lambda \frac{1}{2} \varepsilon + (1 - \lambda) \frac{1}{2} \left(1 - \frac{\delta}{\theta}\right) \varepsilon\]  
\[\hat{test}_{t+L+1} = (1 - \lambda) \frac{1}{2} \varepsilon + \lambda \frac{1}{2} \left(1 - \frac{\delta}{\theta}\right) \varepsilon.\]  

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

\[19\text{See the appendix for a proof of this result.}\]
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 \( \varepsilon \) 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 \( \tau \) but a negative grade level association in the school year starting in \( \tau + L + 1 \) as long as \( \delta \leq \theta \). When \( \delta = \theta \) the positive association is stronger than the negative association but this changes as \( \delta \) 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 school year \( \tau \) come either from birth cohort \( \tau - L \) or birth cohort \( \tau - L - 1 \). Let \( \varepsilon' \) be the shock to the average skills of girls in birth cohort \( \tau - L \) and \( \varepsilon'' \) 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 \( \tau \) 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

\[
E(\tilde{\text{girls}}_{t\tau} \tilde{\text{test}}_{t\tau} | \tau) = \frac{(2\lambda - 1)\delta}{16\theta^2} \text{Var}(\varepsilon)
\]

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 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. 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 schools and birth cohorts.

\[ p'_{fs} = p + \nu_{fs} \quad \text{and} \quad p'_{ms} = p + \nu_{ms} \]

where \( \nu_{fs}, \nu_{ms} \) are shocks with mean zero with \( Var(\nu_{fs}) = Var(\nu_{ms}) = Var(\nu) > 0 \) and \( Correl(\nu_{fs}, \nu_{ms}) = \rho_{\nu} \). Suppose also that the average skills of boys and girls are identical and the same across schools, birth cohorts, and gender, \( \alpha'_{fs} = \alpha'_{ms} = \alpha \), and that there is the same share of boys and girls in each birth cohort, \( \phi'_{t} = 1/2 \).

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 peer effect of girls on girls’ academic achievement and a spurious positive peer effect of girls on boys’ achievement if

\[ (2\lambda - 1) \left( 1 - \frac{\delta}{2\theta} \right) > 0. \]

If \( (2\lambda - 1) \left( 1 - \frac{\delta}{2\theta} \right) < 0 \), the grade level approach would yield a spurious positive peer effect of girls on girls’ academic achievement and a spurious negative peer effect of girls on boys’ achievement. Only if \( (2\lambda - 1) \left( 1 - \frac{\delta}{2\theta} \right) = 0 \) 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 peer effect of girls on girls’ academic achievement when using the grade level approach. The argument for boys is analogous.

2.2.4 Controlling for the Effect of Grade Retention at the Individual Level

It is natural to wonder whether spurious gender peer effects at the grade level could be avoided by controlling for the effect of grade retention on academic achievement at

---

20 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.

21 See the appendix for a proof of this result.
the individual level. Even in our (relatively simple) model based on a uniform skill distribution, this is not the case. Controlling for grade retention at the individual level eliminates spurious gender peer effects due to the effect of gender composition at the birth cohort level on the probability that a HG girl has been retained in the past. But the grade level approach continues to yield spurious gender peer effects when grade level differences in the share of girls are partly driven by shocks to academic skills or to retention thresholds.

To see this, notice that controlling for the effect of grade retention on academic achievement at the individual level amounts to adjusting the academic achievement of retained students by the average gap in academic achievement between retained and non-retained students in the same grade and school. This gap is \( \theta - \delta \), see \((4)\) and \((5)\). Hence, the average test score at the grade level adjusted for the effect of grade retention at the individual level becomes

\[
\text{test}^{\text{REA}}_{gs\tau} = \mu_{gs\tau} E(\text{test}_{igs}^{\tau-L} | \text{non-retained}) + (1 - \mu_{gs\tau})\left( E(\text{test}_{igs}^{\tau-L-1} | \text{retained}) + (\theta - \delta) \right),
\]

which differs from \( \text{test}_{gs\tau} \) in \((8)\) only in the \( \theta - \delta \) term (REA stands for retention-effect adjusted). Following the steps we used to obtain \((16)\) then yields

\[
E(\hat{\text{girlsh}}_{gs\tau}^{\text{REA}} | \tau) = 0.
\]

Hence, once we have controlled for the effect of grade retention on academic achievement at the individual level, shocks to gender composition no longer result in spurious gender peer effects when using a grade level approach. On the other hand, shocks to skills and to academic thresholds continue to result in spurious gender peer effects. This can be derived following the steps that led to \((22)\) and \((24)\) using \( \text{test}^{\text{REA}}_{gs\tau} \) instead of \( \text{test}_{gs\tau} \).

### 2.2.5 Estimating Gender Peer Effects Using Variation Across Classrooms

Another interesting question is whether grade retention can also affect estimates of gender peer effects based on variation in the share of girls across classrooms within a grade and school. To see that it can, suppose that retained students are distributed randomly across classrooms. In this case, classrooms that randomly started out with more girls in first grade will tend to have a lower probability in higher grades that one of the girls in class has been retained in the past. These classrooms will also have a higher probability that one of the boys in class has been retained in the past. Hence, grade retention results in the same effects we discussed in the previous sections. When retained students are not distributed randomly across classrooms, the consequences of grade retention will depend on the system being used.

### 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 assuming, as in the previous section, that the share of girls at the birth cohort level is subject to i.i.d. shocks.

\(^{22}\)See the appendix for a proof of these results.
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 $t$

$$
\alpha_{gs}^t = \tilde{\alpha}_{gs}^t + \pi^g_\phi \phi_s^t
$$

where $\pi^g_\phi$ is the strength of the gender peer effect on the academic skills of students of gender $g$ and $\tilde{\alpha}_{gs}^t$ captures exogenous shocks to skills.\(^{23}\) Gender peer effects could also affect the academic thresholds used to determine who is retained for an extra year in LG. We therefore allow the thresholds for grade retention to depend on the share of girls in birth cohort $t$

$$
\phi_{gs}^t = \tilde{\phi}_{gs}^t + \pi^g_\phi \phi_s^t
$$

where $\pi^g_\phi$ is the strength of the gender peer effect on the threshold for grade retention applied to students of gender $g$ and $\tilde{\phi}_{gs}^t$ captures exogenous shocks. Moreover, there could be gender peer effects in high grade (HG) skill accumulation. To capture this effect we allow mean academic skills accumulated in HG in school year $\tau$ to depend on the share of girls in HG defined in (9)

$$
\omega_{gs\tau} = \tilde{\omega}_{gs\tau} + \pi^g_{\omega_{ gs\tau}} \phi_s^\tau
$$

where $\pi^g_{\omega_{ gs\tau}}$ is the strength of the gender peer effect and $\tilde{\omega}_{gs\tau}$ captures exogenous shocks.

Suppose now that we have data on the test performance in HG of students from birth cohorts $t$ and $t-1$ for a large number of schools. 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 yields the same asymptotic result as 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, $\phi_s^t - \phi_s^{t-1}$ (across schools). 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

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

where $\lambda_{gs}^t$ is the share of non-retained students of gender $g$ in birth cohort $t$ and school

\(^{23}\)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.
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

$$
\beta^B_g = \pi^\alpha_g + \pi^\omega_g \left( \lambda^2 + (1-\lambda)^2 + \lambda \frac{\partial (\lambda_f - \lambda_m)}{\partial \phi} \right) - \delta \frac{\partial \lambda_{gs}}{\partial \phi}
$$

where we have linearized (29); $\lambda_g = (\bar{\alpha} + \pi^\alpha_g \phi + \theta - \bar{\pi} - \pi^\omega_g \phi)/2\theta$ combines (2) and (25)-(26); and $\partial \lambda_{gs}/\partial \phi = (\pi^\alpha_g - \pi^\omega_g)/29$. 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 $\pi^\omega_g$ 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.

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 ($\pi^p_g \neq 0$ but $\pi^\alpha_g = \pi^\omega_g = 0$ 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 $\delta$, 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 $\pi^\alpha_g = \pi^\omega_g = \pi^p_g = 0$ combined with (2) and (25)-(27) implies $\beta^B_g = 0$.\footnote{We linearize around $\tilde{\omega}_{gs} = \tilde{\omega}$, $\tilde{p}_{gs} = \tilde{p}$, $\tilde{\alpha}_{gs} = \alpha$, $\delta_{gs} = \delta$, and $\phi_t = 1/2$.}

\footnote{To see this, note that $\lambda$ 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 $\lambda$ 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 $\lambda(1-\lambda)(1-\lambda)$. 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.}
2.3.1 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 $\tau$ instrumented by the share of girls in birth cohort $\tau - L$. If shocks to the share of girls at the birth cohort level (the instrument in this approach) are independent of shocks to skills at the birth cohort level and of shocks to academic thresholds, this approach eliminates the spurious gender peer effects due to skill and threshold shocks described in sections 2.2.2 and 2.2.3. But the approach would continue to yield spurious gender peer effects. This is because shocks to the share of girls at the birth cohort level (the instrument) affect the probability that a girl/boy in HG has been retained in the past as explained in section 2.2.1 and hence the skill distribution at the grade level. As a result, the instrument affects individual test performance even if there are no true gender peer effects.

However, it is possible to avoid spurious gender peer effects when using the grade level approach by simultaneously controlling for grade retention at the individual level, as in section 2.2.4, and instrumenting the share of girls in HG in the school year starting in $\tau$ by the share of girls in birth cohort $\tau - L$. This approach would yield a strength of gender peer effects $\beta_{g}^{REA,IV} = (1/2)(\pi_{g}^{a} + \pi_{g}^{p})/[1 + (1/4\lambda)\partial(\lambda_{f} - \lambda_{m})/\partial\phi] + \pi_{g}^{w}$ where the parameters are defined in the previous section. This effect can be smaller or larger than the strength of gender peer effects obtained with the birth cohort approach in (30). This is because the two approaches look at different gender peer effects as far as retained students are concerned. The birth cohort approach relates the academic achievement of retained students to the gender composition of the peers they started out with in lower grades. The instrumental-variables approach relates the academic achievement of retained students to the gender composition of their peers after grade retention.

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. 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

---

26 The proof of this result is in the appendix. Note that $\beta_{g}^{REA,IV} = 0$ if $\pi_{g}^{a} = \pi_{g}^{w} = \pi_{g}^{p} = 0$, which implies that this approach does not generate spurious gender peer effects.

27 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).
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). Spanish primary schools permit grade retention and rates of grade retention 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.

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. 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 \( a_{i\tau} = \frac{z_{i\tau} - \mu_\tau}{\sigma_\tau} \), where \( z_{i\tau} \) is the raw score of test taker \( i \) in the school year starting in year \( \tau \), and \( \mu_\tau \) and \( \sigma_\tau \) are the mean

---

28This 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 Spain the data come from the national statistical institute’s INEbase (2013).

29Throughout 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.

30These children can be thought of as being retained in kindergarten.
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.\footnote{Interestingly, Lavy and Schlosser (2011) find that gender peer effects are statistically insignificant in Hebrew, which is probably closest to reading in our data.}

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 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. 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. 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.

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. 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

---

32 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.

33 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.

34 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.
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.\textsuperscript{35} 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.\textsuperscript{36}

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

\begin{equation}
\begin{aligned}
x_{igs}^t &= \gamma_g^s + \gamma_g^t + \gamma_g^f Birthcohortschool_{s}^t + u_{igs}^t \\
\end{aligned}
\end{equation}

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 \(u_{igs}^t\) is the residual. School level differences and gender specific trends in characteristic \(j\) for gender \(g\) are captured by \(\gamma_g^s\) and \(\gamma_g^t\), respectively. The parameter of interest is \(\gamma_g^f\) which captures whether student characteristic \(j\) varies with the share of girls in the birth cohort.

Table 1 contains our estimates of \(\gamma_g^f\) in (31) as well as robust standard errors clustered by school. We also report a statistic for the hypothesis that \(\gamma_g^f = ... = \gamma_g^J = 0\) 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.\textsuperscript{37} The results in table 1 indicate that student characteristics are balanced with respect to the share of girls in a birth cohort.\textsuperscript{38}

\textsuperscript{35} 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 incentivizes parents to pick a neighborhood school (e.g. Caldera and Andrews, 2011; Calsamiglia and Guell, 2013).

\textsuperscript{36} 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).

\textsuperscript{37} We also get balancedness at the birth cohort level in the other samples, see the results in table 2.

\textsuperscript{38} 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.
the 74 characteristics we examine for boys and girls, only 4 yield statistically significant \( \gamma_j \) at the 5 or 10 percent level (the number one would expect based on type I error). Moreover, the hypothesis \( \gamma_1^g = \ldots = \gamma_J^g = 0 \) 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.\(^{39}\) 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 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

\[
\text{test}_{i gs}^t = \alpha_{gs} + \alpha_{tg} + \sum_j \alpha_j^g x_{ij}^t (1 - D_{igt}) + \sum_j \gamma_j^g D_{igt} + v_{igt} + \beta_g \text{BirthcohortGirlshare}_{iga}^t + v_{iga}^t
\]

where \( \text{test}_{i gs}^t \) 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 \( v_{iga}^t \) is the residual. The table reports the least squares estimates of \( \beta_g \) with robust standard errors clustered by school. We report results for the average test score and the test score in mathematics.

\(^{39}\) The regressions are \( \text{test}_{i gs}^t = \alpha_{gs} + \alpha_{tg} + \sum_j \alpha_j^g x_{ij}^t (1 - D_{igt}) + \sum_j \gamma_j^g D_{igt} + v_{igt} + \beta_g \text{BirthcohortGirlshare}_{iga}^t + v_{iga}^t \) where \( \text{test}_{i gs}^t \) refers to standard test scores, \( i \) to individuals, \( t = 2003, 2004 \) to birth cohorts, \( g \) to gender, \( s \) to schools, \( D_{igt} \) 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 \( v_{iga}^t \) is the residual. Individual characteristics account for around 11 to 15 percent of the variance in test performance.
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). 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 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 indicate 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.

\[^{40}\text{The median number of students in schools with at most two classes per grade is 36 and the maximum number of students 58. The median number of students in schools with more than two classes per grade is 73 and the maximum number 152.}\]
of a standard deviation.  

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

\[ test_{tgst} = \alpha_g + \alpha_{g\tau} + \alpha X_{tgst} + \beta_g \text{GradeGirlshare}_{gst} + \nu_{gst} \]  

(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 point in the opposite 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. 

Table 5 reestimates the specifications in table 4 after adding a control indicating whether the student has repeated a grade (an indicator variable that is unity if and only if the student has repeated a grade). This does not affect the main conclusions.

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 (and therefore negatively selected) 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 6 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 6 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.

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

\[ \text{We do not include the school year starting in 2010 to keep the empirical analysis using the grade level approach as close as possible to that using the birth cohort approach. When we include the 2010 school year, the results are very similar to those in table 4 except that estimates become more precise and that the (positive) effect of girls on girls' academic achievement in schools with one class per grade is often statistically significant. It is interesting to note that the effect of the share of girls in a grade on the test performance of girls in table 4 is statistically significant in the sample with all schools, while the birth cohort approach yielded statistically insignificant effects for all schools. As already mentioned, if gender peer effects are at the classroom level, the results using the birth cohort approach can be explained by the share of girls at the birth cohort level being a noisy measure of the share of girls at the classroom level in schools with several classes per grade. On the other hand, the grade level approach will produce the spurious gender peer effects described in section 2.2 no matter how many classes there are per grade as long as grade level differences in gender composition are mostly driven by shocks to gender composition or skills at the birth cohort level.} \]
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.\textsuperscript{13}

5 Counterfactual Analysis

As rates of grade retention in Spain are higher than 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 $a_{tgs}$ drawn from a school, gender, and birth cohort specific distribution $F_{tgs}$. Students with skills $a_{tgs}$ above the academic threshold $p_{tgs}$ are promoted to high grade (HG). Students with skills below the academic threshold spend one extra year in LG before being admitted to HG.\textsuperscript{44} After one year in HG, students leave primary school.

In the model in section 2 we took the distribution $F_{tgs}$ 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, $\omega_{tgs}^T = \omega_{gs}$.\textsuperscript{45} This implies that the HG test scores of students in birth cohort

\textsuperscript{13}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).

\textsuperscript{44}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) INElbase 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.

\textsuperscript{45}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.
t who have not been retained during primary school are $a_{tgs} + \omega_{gs}$ with $a_{tgs}$ drawn from a distribution $F_{tgs}$ truncated below at $p_{tgs}$ and that the HG test scores of students in birth cohort $t$ who have been retained during primary school are $a_{tgs} + \omega_{gs} + \delta_{tgs}$ with $a_{tgs}$ drawn from $F_{tgs}$ truncated above at $p_{tgs}$. This is equivalent to assuming that the HG test scores of non-retained students in birth cohort $t$ are equal to $a_{tgs}$ drawn from $F_{tgs}$ shifted by $\omega_{gs}$ and truncated below at $p_{tgs}$ and that the HG test scores of retained students are equal to $a_{tgs} + \delta_{tgs}$ with $a_{tgs}$ drawn from $F_{tgs}$ shifted by $\omega_{gs}$ and truncated above at $p_{tgs} + \omega_{gs}$. We can therefore simplify the notation by absorbing HG skill accumulation into the mean of $F_{tgs}$ and the academic thresholds for grade retention $p_{tgs}$ 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 $\phi_{gs}$ 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 $F_{tgs}$; (b) the academic thresholds for grade retention $p_{tgs}$; (c) the changes $\delta_{tgs}$ in skills that come with grade retention; and (d) the probability that a student is female $\phi_{gs}$. 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 \(\alpha_{tg}^t \sim U(\alpha_{tg}^t - \theta, \alpha_{tg}^t + \theta)\). The distribution parameters we need to calibrate are the school, birth cohort, and gender specific means \(\alpha_{tg}^t\) and the parameter \(\theta\) governing skill dispersion.

Calibrating the Parameters of the 2003 and 2004 Birth Cohorts We calibrate the school, birth cohort, and gender specific means \(\alpha_{tg}^t\) and the parameter \(\theta\) governing the dispersion of skills jointly with the academic thresholds for grade retention \(p_{tg}^t\) 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, \(\lambda_{tg}^t\); (ii) the average test score of non-retained students at the school, birth cohort, and gender level, \(E_{tg}^t(\text{test} | \text{non-retained})\); and (iii) the variance of the test score of non-retained students, \(\text{Var}_{tg}^t(\text{test} | \text{non-retained})\), averaged across schools, birth cohorts, and gender. Denoting the cumulative distribution of the uniform \(U(\alpha_{tg}^t - \theta, \alpha_{tg}^t + \theta)\) by \(F_{tg}^t\), this yields the first three calibration equations

\[
1 - F_{tg}^t(p_{tg}^t) = \lambda_{tg}^t
\]

\[
E_{tg}^t(a | a \geq p_{tg}^t) = E_{tg}^t(\text{test} | \text{non-retained})
\]

\[
\frac{1}{4S} \sum_{t,s,g} \text{Var}_{tg}^t(a | a \geq p_{tg}^t) = \frac{1}{4S} \sum_{t,s,g} \text{Var}_{tg}^t(\text{test} | \text{non-retained})
\]

where \(S\) is the number of schools. The changes \(\delta_{tg}^t\) 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, \(E_{tg}^t(\text{test} | \text{retained})\),

\[
\delta_{tg}^t + E_{tg}^t(a | a < p_{tg}^t) = E_{tg}^t(\text{test} | \text{retained})\]

\[\text{46}\]

Our calibration yields positive average and median changes \(\delta_{tg}^t\) across 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 and Lefgren (2004) for the US suggest a positive short-term effect on skills of grade
This yields a system of $12S + 1$ calibration equations with $12S + 1$ parameters, which can be solved in closed form.47

The probability $\phi^t_s$ 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 $\theta$ 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 $\alpha_{fs}^{2002}$ and $\alpha_{ms}^{2002}$ 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 $\delta_{fs}^{2002}$ and $\delta_{ms}^{2002}$ 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 $\alpha_{gs}^t$ and standard deviation $\theta$, $a_{gs}^t \sim N(\alpha_{gs}^t, \theta^2)$. 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.48

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. 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.

47 Using the formula for the variance of uniform distributions and (2), (34), and (36) yields that $\theta$ is the positive root of $3 \sum_{t,s,g} \text{Var}_{gs} \sqrt{\sum_{t,s,g} (\lambda_{gs})^2}$. Given $\theta$, (2) and (4) imply that (34) and (35) are linear in $\alpha_{gs}^t$ and $p_{gs}$.

48 While these and subsequent numerical solutions could in principle yield multiple solutions, the solutions found appeared to be unique in practice.
We now present a first approach that accounts for these bounds on raw test scores.\footnote{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.}

To do so we assume that students’ raw skills $z_{t}^t$ are generated according to

$$z_{t}^t = 10 \left( \frac{1}{1 + \exp(-v_{tgs}^t)} \right) \quad \text{with } v_{tgs}^t \sim N(\alpha_{tgs}^t, \theta^2)$$

which ensures $0 \leq z_{t}^t \leq 10$. We use (38) to obtain the density and cumulative distribution of students’ standard test scores as $a_{t}^t = (z_{t}^t - \mu_t^t)/\sigma_t^t$ where $\mu_t^t$ and $\sigma_t^t$ are the mean and standard deviation of raw scores in the year non-retained students from birth cohort $t$ took the test. The parameters $\alpha_{tgs}^t$ in (38) as well as $p_{tgs}^t$ and $\theta$ 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 $\alpha_{tgs}^t$, $p_{tgs}^t$, and $\theta$ for the 2003 and 2004 birth cohorts, we obtain the values of $\delta_{t}^t$ 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 $z_{t}^t \sim T(l_{tgs}^t, u_{tgs}^t, m_{tgs}^t)$ where $l$ and $u$ are the lower and upper bound and $m \in [l, u]$ 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 $a_{t}^t = (z_{t}^t - \mu_t^t)/\sigma_t^t$ where $\mu_t^t$ and $\sigma_t^t$ 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 $u = 10$. 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 $m_{tgs}^t$ and the academic threshold for grade retention $p_{tgs}^t$).

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.\footnote{It is straightforward to determine which of the two assumptions has to be relaxed to find a solution.} Once we have obtained a solution...
for the parameters of the triangular skill distribution and the academic thresholds, we calibrate the values of $\delta_{gs}$ 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 $\phi_{tgs}$. Once the gender of students has been determined, we draw their skills $a_{tigs}$ from the school, birth cohort, and gender specific skill distributions. Students with skills $a_{tigs}$ above the threshold for grade retention $p_{tgs}$ are assumed to enter sixth grade in year $t + 5$ and achieve a test score $a_{tigs}$ in the standardized test. Students with skills $a_{tigs}$ below the threshold for grade retention $p_{tgs}$ are assumed to enter sixth grade in year $t + 6$ and achieve a test score $a_{tigs} + \delta_{gs}$. 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

$$test_{gs\tau} = \mu_{gs\tau}E(test_{tigs}^{\tau-5} | non-retained) + (1 - \mu_{gs\tau})E(test_{tigs}^{\tau-6} | retained)$$

where $\mu_{gs\tau}$ is the simulated share of non-retained girls or boys among sixth graders of the same gender, while $E(test_{tigs}^{\tau-5} | non-retained)$ and $E(test_{tigs}^{\tau-6} | retained)$ 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 $\tau$. 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

$$test_{tgs} = \lambda_{gs}^{t} E(test_{tigs} | non-retained) + (1 - \lambda_{gs}^{t})E(test_{tigs} | retained)$$

where $\lambda_{gs}^{t}$ is the simulated share of non-retained girls or boys among students of the same gender in birth cohort $t$, while $E(test_{tigs} | non-retained)$ and $E(test_{tigs} | retained)$ 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, $test_{tgs}^{2004} − test_{tgs}^{2003}$, 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 panel A and B of table 7. We report point estimates and standard errors averaged across 100 simulations.\(^{51}\) 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 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).\(^{52}\) 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

\(^{51}\) As we have 1000 students per school and birth cohort, results vary little across simulations.

\(^{52}\) 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.
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 table 7. 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.

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 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 table 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 table 7 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. 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 peer effects when girls and boys learn together. The selection of students with different skills into different schools can be bypassed by

53 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.

54 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 \( \pi^g = 0 \), 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 \( \pi^g > 0 \), 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 \( \pi^g = 0 \) as our data do not allow us to calibrate the distribution of academic skills at different stages of primary school.
exploiting within-school variation in gender composition as in Hoxby (2000) and Lavy and Schlosser (2011) or random assignment to schools as in Lee et al. (2014). But what about selection within schools due to grade retention of students who have not acquired certain academic skills or achieved learning goals? We have argued that grade retention generally leads to spurious gender peer effects when estimation is based on grade level variation in the share of girls. This is because any initial randomness in the distribution of girls and boys — at entry into primary or secondary school for example — is undone later by grade retention of low-skill students.

As grade retention is practiced in most school systems, we proposed estimating peer effects when girls and boys learn together by exploiting within-school differences in gender composition across birth cohorts. The birth cohort approach is an intention-to-treat approach examining 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. Another feature of the birth cohort approach is that it is equivalent to the grade level approach of Hoxby (2000) and Lavy and Schlosser (2011) when grade retention is not permitted and students start school according to the enrollment rule. We then used the birth cohort approach to estimate gender peer effects in Spanish primary schools, which practice grade retention. We found a statistically significant positive peer effect of girls on the individual academic achievement of boys by the end of primary school, but a statistically insignificant effect of girls on girls’ academic achievement.
Appendix

The Grade Level Approach To prove the results in sections 2.2.1-2.2.3, we first linearize the within-school grade level change in the share of girls in high grade (HG), \( \Delta \text{girlsh}_{st} = \text{girlsh}_{st} - \text{girlsh}_{s,t-1} \), around \( \phi_s^t = 1/2, \alpha_{gs}^t = \alpha, \) and \( p_{gs}^t = p \). Making use of (2) and (9), this yields

\[
(A1) \quad \Delta \text{girlsh}_{st} = (1/4)(\Delta \lambda_{fs}^{\tau-L} - \Delta \lambda_{ms}^{\tau-L} - \Delta \lambda_{fs}^{\tau-L-1} + \Delta \lambda_{ms}^{\tau-L-1})
+ \lambda \Delta \phi_s^{\tau-L} + (1 - \lambda) \Delta \phi_s^{\tau-L-1}
\]

where \( \lambda = (\alpha + \theta - p)/2\theta \), \( \Delta \lambda_{gs}^t = (\alpha_{gs}^t - p_{gs}^t)^t - (\alpha_{gs}^{t-1} + p_{gs}^{t-1})/2\theta \), 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, \( \Delta \text{test}_{gst} = \text{test}_{gst} - \text{test}_{gst,s,t-1} \), using (4)-(9) yields

\[
(A2) \quad \Delta \text{test}_{gst} = (1/2) \lambda (\Delta \alpha_{gs}^{\tau-L} + \Delta p_{gs}^{\tau-L-1}) + (1/2)(1 - \lambda)(\Delta \alpha_{gs}^{\tau-L-1} + \Delta p_{gs}^{\tau-L-1})
+ (\theta - \delta)(2(1 - \lambda)\lambda(2\phi_s^{\tau-L-1} - \phi_s^{\tau-L} - \phi_s^{\tau-L-2}) + (1 - \lambda)\Delta \alpha_{gs}^{\tau-L} + \lambda \Delta \lambda_{gs}^{\tau-L-1})
\]

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 \( \Delta \text{girlsh}_{st} \) and \( \Delta \text{test}_{gst} \) 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 \( \Delta \text{girlsh}_{gst} = \lambda \phi_s^{\tau-L} - \lambda \phi_s^{\tau-L-1} + (1 - \lambda)\phi_s^{\tau-L-1} - (1 - \lambda)\phi_s^{\tau-L-2} \) and (A2) to \( \Delta \text{test}_{gst} = 2(\theta - \delta)(1 - \lambda)\lambda(2\phi_s^{\tau-L-1} - \phi_s^{\tau-L} - \phi_s^{\tau-L-2}) \). Hence, the assumption of i.i.d. shocks to the share of girls in a birth cohort implies \( \text{Cov}(\Delta \text{girlsh}_{gst}, \Delta \text{test}_{fst}|\tau) = 6(\theta - \delta)(2\lambda - 1)(1 - \lambda)\text{Var}(\eta) \). 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 \( (\theta - \delta)(2\lambda - 1)\text{Var}(\eta) > 0 \), 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 \( \text{Cov}(\Delta \text{girlsh}_{gst}, \Delta \text{test}_{gst}|\tau) = 3(2\lambda - 1)\delta(1 - \rho_c)\text{Var}(\varepsilon)/16\theta^2 \). As \( \theta > 0 \), this implies a strictly positive least squares slope when regressing (A2) on (A1) across schools if \( (2\lambda - 1)\delta(1 - \rho_c)\text{Var}(\varepsilon) > 0 \), which proves (18). The result in (24) for shocks to the academic shocks used for grade retention can be proven analogously.

Controlling for the Effect of Grade Retention at the Individual Level We now prove the result in section 2.2.4. To do so, we linearize \( \Delta \text{test}_{gst}^{\text{RE}} \) following the same steps as in the previous section, where \( \text{test}_{gst}^{\text{RE}} \) is defined in section 2.2.4 as

\[
(A3) \quad \text{test}_{gst}^{\text{RE}} = \mu_{gst} E(\text{test}_{igt}^{\tau-L}|\text{non-retained})
+ (1 - \mu_{gst})(E(\text{test}_{igt}^{\tau-L-1}|\text{retained}) + (\theta - \delta))
\]
this yields

\[
(\text{A4}) \quad \Delta \text{test}_{gs}^{\text{REA}} = (1/2)\lambda \left( \Delta \alpha_{gs}^{-L} + \Delta \rho_{gs}^{-L-1} \right) \\
+ (1/2)(1 - \lambda)(\Delta \alpha_{gs}^{-L-1} + \Delta \rho_{gs}^{-L-1}).
\]

Following the same steps as in the previous section, it can now be shown that if the grade level variation in the share of girls is driven by skill shocks at the birth cohort level then a least squares regression of the retention-effect adjusted average test score at the grade level in (A4) on the grade level change in the share of girls in HG in (A2) yields a strictly positive slope if and only if \((2\lambda - 1)(1 - \rho_\lambda)\text{Var}(\varepsilon) > 0\). On the other hand, if the grade level variation in the share of girls is driven by shocks to the academic thresholds used for grade retention, a least squares regression of the retention-effect adjusted average test score at the grade level in (A4) on the grade level change in the share of girls in HG in (A2) yields a strictly positive slope if and only if \((1 - 2\lambda)(1 - \rho_\eta)\text{Var}(\eta) > 0\).

Note also that (A4) – in contrast to (A2) – does not depend on the share of girls at the birth cohort level \((\phi)\), which implies \(E(\Delta \text{girls}_{gs}\Delta \text{test}_{gs}^{\text{REA}}) = 0\) if the grade level variation in the share of girls is driven solely by shocks to the gender composition at the birth cohort level. Hence, adjusting for the effects of grade retention at the individual level eliminates spurious gender peer effects due to shocks to gender composition at the birth cohort level.

A Grade Level Instrumental-Variables Approach  We now prove the result in section 2.3.1. To do so, we need to take into account that in the model with peer effects in section 2.3, the retention-effect adjusted average test score at the grade level becomes \(\text{test}_{gs}^{\text{REA}} + \omega_{gs}\) where \(\text{test}_{gs}^{\text{REA}}\) is defined in (A3) and \(\omega_{gs}\) in (27) captures learning in HG (and possible gender peer effects in HG). Linearizing as in the previous section we obtain

\[
(\text{A5}) \quad \Delta \text{test}_{gs}^{\text{REA}} + \Delta \omega_{gs} = (1/2)\lambda \left( \Delta \alpha_{gs}^{-L} + \Delta \rho_{gs}^{-L-1} \right) \\
+ (1/2)(1 - \lambda)(\Delta \alpha_{gs}^{-L-1} + \Delta \rho_{gs}^{-L-1}) + \Delta \omega_{gs}.
\]

To derive the result of the instrumental-variables approach described in section 2.3.1, it is easiest to think of the instrumental-variables estimate as the ratio of the reduced-form effects of the share of girls at the birth cohort level \((i)\) on the retention-effect adjusted average test score at the grade level in (A5) and (ii) on the share of girls at the grade level in (A1). These effects are, respectively,

\[
\frac{\partial (\Delta \text{test}_{gs}^{\text{REA}})}{\partial (\Delta \phi_{gs}^{-L})} = (1/2)\lambda(\pi_\alpha' + \pi_\rho') + \pi_\omega' \left( \lambda + (1/4)\frac{\partial(\lambda_f - \lambda_m)}{\partial \phi} \right)
\]

\[
(\text{A6})
\]

\[
\frac{\partial (\Delta \text{girls}_{gs})}{\partial (\Delta \phi_{gs}^{-L})} = \lambda + (1/4)\frac{\partial(\lambda_f - \lambda_m)}{\partial \phi}
\]

\[
(\text{A7})
\]

where we used (25)-(27). The ratio of the reduced-form effects in (A6) and (A7) is

\[
\beta_{gs}^{\text{REA}, IV} = \frac{(1/2)(\pi_\alpha' + \pi_\rho')}{1 + (1/4)\frac{\partial(\lambda_f - \lambda_m)}{\partial \phi}} + \pi_\omega'.
\]

38
References


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


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

Figure 1: A School System with Grade Retention

<table>
<thead>
<tr>
<th>Birth cohort $t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starts primary school in year $t$</td>
</tr>
</tbody>
</table>

1)  
2)  
...  
$L$)  

$L$ years in lower grades  

extra year in lower grades  

skills change by $\delta^{t}_{gs}$  

- non-retained students birth cohort $t$ score  
  $a^{t}_{gs} + \omega_{gs,t+L}$  

- retained students birth cohort $t$ score  
  $a^{t}_{gs} + \delta^{t}_{gs} + \omega_{gs,t+L+1}$  

Take the test  

one year in high grade  
leave primary school
Table 1: Detailed Balancedness at the Birth Cohort Level (Schools with at Most Two Classes per Grade)

<table>
<thead>
<tr>
<th>Father profession</th>
<th>Boys</th>
<th>Girls</th>
<th>Mother profession</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession 1</td>
<td>0.003</td>
<td>0.012</td>
<td>Profession 1</td>
<td>0.002</td>
<td>-0.008</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.015)</td>
<td></td>
<td>(0.006)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>Profession 2</td>
<td>0.011</td>
<td>-0.025</td>
<td>Profession 2</td>
<td>0.058</td>
<td>-0.035</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.038)</td>
<td></td>
<td>(0.037)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>Profession 3</td>
<td>0.079*</td>
<td>-0.001</td>
<td>Profession 3</td>
<td>-0.019</td>
<td>-0.040</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.043)</td>
<td></td>
<td>(0.044)</td>
<td>(0.043)</td>
</tr>
<tr>
<td>Profession 4</td>
<td>-0.025</td>
<td>0.002</td>
<td>Profession 4</td>
<td>0.038</td>
<td>-0.011</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td></td>
<td>(0.030)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>Profession 5</td>
<td>-0.059</td>
<td>-0.023</td>
<td>Profession 5</td>
<td>-0.030</td>
<td>0.113**</td>
</tr>
<tr>
<td></td>
<td>(0.043)</td>
<td>(0.046)</td>
<td></td>
<td>(0.043)</td>
<td>(0.046)</td>
</tr>
<tr>
<td>Profession 6</td>
<td>-0.065</td>
<td>0.034</td>
<td>Profession 6</td>
<td>0.000</td>
<td>0.017*</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.046)</td>
<td></td>
<td>(0.010)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Profession 7</td>
<td>0.018</td>
<td>-0.017</td>
<td>Profession 7</td>
<td>0.004</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.029)</td>
<td></td>
<td>(0.019)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>Profession 8</td>
<td>0.038</td>
<td>0.027</td>
<td>Profession 8</td>
<td>-0.054</td>
<td>-0.033</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.029)</td>
<td></td>
<td>(0.053)</td>
<td>(0.052)</td>
</tr>
</tbody>
</table>

| Father education  | No degree  | -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 | Mother | -0.029 | Before 3 | -0.023 | -0.023 |
|                       |       | (0.046) | (0.047) | (0.060) | (0.057) |
|                       | Father | -0.073 | -0.020 | 0.031 | 0.030 |
|                       |       | (0.082) | (0.097) | (0.056) | (0.056) |
|                       | One sibling | 0.130 | -0.173 | 0.011 | -0.007 |
|                       |       | (0.110) | (0.118) | (0.019) | (0.018) |
|                       | More siblings | -0.126 | 0.077 | -0.008 | 0.000 |
|                       |       | (0.084) | (0.086) | (0.013) | (0.011) |
|                       | Other family | -0.023 | 0.021 | 0.040 | 0.022 |
|                       |       | (0.068) | (0.064) | (0.047) | (0.042) |
|                       | Other | 0.044 | -0.023 | 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 with just over 60000 students in total. *, **, *** 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

<table>
<thead>
<tr>
<th></th>
<th>Panel A: Two-Class Schools</th>
<th>Panel B: One-Class Schools</th>
<th>Panel C: All Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
<td>Boys</td>
</tr>
<tr>
<td>Average score</td>
<td>-0.007</td>
<td>0.007</td>
<td>-0.020</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.038)</td>
<td>(0.054)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.005</td>
<td>0.013</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.033)</td>
<td>(0.045)</td>
</tr>
</tbody>
</table>

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 with just over 95000 students in total; 908 schools with at most two classes per grade with just over 60000 students in total; and 331 schools with one class per grade with just over 13000 students in total. *, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.
Table 3: Gender Peer Effects Using the Birth Cohort Approach

<table>
<thead>
<tr>
<th>Panel A: All Schools</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
<td>Boys</td>
<td>Girls</td>
<td></td>
</tr>
<tr>
<td>Average score</td>
<td>0.150</td>
<td>0.085</td>
<td>0.155</td>
<td>0.085</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>(0.106)</td>
<td>(0.110)</td>
<td>(0.102)</td>
<td>(0.105)</td>
<td>(0.103)</td>
</tr>
<tr>
<td>Math score</td>
<td>0.160</td>
<td>0.090</td>
<td>0.17*</td>
<td>0.090</td>
<td>0.17*</td>
</tr>
<tr>
<td></td>
<td>(0.104)</td>
<td>(0.111)</td>
<td>(0.102)</td>
<td>(0.109)</td>
<td>(0.103)</td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Two-Class Schools</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
<td>Boys</td>
<td>Girls</td>
<td></td>
</tr>
<tr>
<td>Average score</td>
<td>0.238**</td>
<td>0.099</td>
<td>0.255**</td>
<td>0.113</td>
<td>0.253**</td>
</tr>
<tr>
<td></td>
<td>(0.115)</td>
<td>(0.122)</td>
<td>(0.112)</td>
<td>(0.117)</td>
<td>(0.113)</td>
</tr>
<tr>
<td>Math score</td>
<td>0.241**</td>
<td>0.113</td>
<td>0.256**</td>
<td>0.128</td>
<td>0.255**</td>
</tr>
<tr>
<td></td>
<td>(0.114)</td>
<td>(0.126)</td>
<td>(0.112)</td>
<td>(0.123)</td>
<td>(0.114)</td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: One-Class Schools</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
<td>Boys</td>
<td>Girls</td>
<td></td>
</tr>
<tr>
<td>Average score</td>
<td>0.263</td>
<td>-0.049</td>
<td>0.313*</td>
<td>0.017</td>
<td>0.317**</td>
</tr>
<tr>
<td></td>
<td>(0.183)</td>
<td>(0.169)</td>
<td>(0.182)</td>
<td>(0.167)</td>
<td>(0.180)</td>
</tr>
<tr>
<td>Math score</td>
<td>0.370**</td>
<td>-0.103</td>
<td>0.397**</td>
<td>-0.037</td>
<td>0.388**</td>
</tr>
<tr>
<td></td>
<td>(0.179)</td>
<td>(0.183)</td>
<td>(0.179)</td>
<td>(0.179)</td>
<td>(0.184)</td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

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 with just over 95000 students in total; 908 schools with at most two classes per grade with just over 60000 students in total; and 331 schools with one class per grade with just over 13000 students in total. *, **, *** 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

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
<th>Boys</th>
<th>Girls</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score</td>
<td>-0.088</td>
<td>0.315***</td>
<td>-0.061</td>
<td>0.288***</td>
<td>-0.047</td>
<td>0.274**</td>
</tr>
<tr>
<td></td>
<td>(0.121)</td>
<td>(0.121)</td>
<td>(0.118)</td>
<td>(0.118)</td>
<td>(0.118)</td>
<td>(0.118)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.057</td>
<td>0.299**</td>
<td>-0.035</td>
<td>0.274**</td>
<td>-0.023</td>
<td>0.260**</td>
</tr>
<tr>
<td></td>
<td>(0.118)</td>
<td>(0.126)</td>
<td>(0.115)</td>
<td>(0.123)</td>
<td>(0.115)</td>
<td>(0.124)</td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Panel B: Two-Class Schools

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
<th>Boys</th>
<th>Girls</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score</td>
<td>0.058</td>
<td>0.406***</td>
<td>0.091</td>
<td>0.381***</td>
<td>0.101</td>
<td>0.355***</td>
</tr>
<tr>
<td></td>
<td>(0.135)</td>
<td>(0.134)</td>
<td>(0.130)</td>
<td>(0.131)</td>
<td>(0.129)</td>
<td>(0.130)</td>
</tr>
<tr>
<td>Math score</td>
<td>0.030</td>
<td>0.349**</td>
<td>0.057</td>
<td>0.326**</td>
<td>0.067</td>
<td>0.295**</td>
</tr>
<tr>
<td></td>
<td>(0.130)</td>
<td>(0.142)</td>
<td>(0.126)</td>
<td>(0.140)</td>
<td>(0.126)</td>
<td>(0.140)</td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Panel C: One-Class Schools

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
<th>Boys</th>
<th>Girls</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score</td>
<td>-0.042</td>
<td>0.222</td>
<td>0.076</td>
<td>0.267</td>
<td>0.046</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>(0.201)</td>
<td>(0.195)</td>
<td>(0.198)</td>
<td>(0.189)</td>
<td>(0.199)</td>
<td>(0.185)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.081</td>
<td>0.121</td>
<td>0.011</td>
<td>0.160</td>
<td>-0.004</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>(0.202)</td>
<td>(0.214)</td>
<td>(0.198)</td>
<td>(0.207)</td>
<td>(0.199)</td>
<td>(0.203)</td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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 with just over 95000 students in total; 908 schools with at most two classes per grade with just over 60000 students in total; and 331 schools with one class per grade with just over 13000 students in total. *, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.
Table 5: Gender Peer Effects Using the Grade Level Approach with Individual Control for Grade Retention

<table>
<thead>
<tr>
<th></th>
<th>Panel A: All Schools</th>
<th>Panel B: Two-Class Schools</th>
<th>Panel C: One-Class Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
<td>Boys</td>
</tr>
<tr>
<td>Average score</td>
<td>-0.101</td>
<td>0.269**</td>
<td>-0.081</td>
</tr>
<tr>
<td></td>
<td>(0.118)</td>
<td>(0.119)</td>
<td>(0.116)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.068</td>
<td>0.252**</td>
<td>-0.052</td>
</tr>
<tr>
<td></td>
<td>(0.116)</td>
<td>(0.126)</td>
<td>(0.114)</td>
</tr>
<tr>
<td>Repeater control</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average score</td>
<td>0.027</td>
<td>0.345***</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>(0.130)</td>
<td>(0.132)</td>
<td>(0.127)</td>
</tr>
<tr>
<td>Math score</td>
<td>0.005</td>
<td>0.298**</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>(0.126)</td>
<td>(0.142)</td>
<td>(0.124)</td>
</tr>
<tr>
<td>Repeater control</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average score</td>
<td>0.012</td>
<td>0.193</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td>(0.195)</td>
<td>(0.189)</td>
<td>(0.193)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.036</td>
<td>0.097</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>(0.197)</td>
<td>(0.211)</td>
<td>(0.195)</td>
</tr>
<tr>
<td>Repeater control</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Individual controls</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Peer group controls</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

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. The repeater control is an indicator variable that is unity if and only if the student has repeated a grade. 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 with just over 95000 students in total; 908 schools with at most two classes per grade with just over 60000 students in total; and 331 schools with one class per grade with just over 13000 students in total. *, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.
Table 6: (Non-)Balancedness of Predicted Test Scores at the Grade Level

<table>
<thead>
<tr>
<th>Panel A: Two-Class Schools</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
</tr>
<tr>
<td>Average score</td>
<td>-0.051</td>
<td>0.058*</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.042</td>
<td>0.059**</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.030)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: One-Class Schools</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
</tr>
<tr>
<td>Average score</td>
<td>-0.098**</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
<td>(0.050)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.087*</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>(0.050)</td>
<td>(0.049)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: All Schools</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Girls</td>
</tr>
<tr>
<td>Average score</td>
<td>-0.041</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>(0.050)</td>
<td>(0.047)</td>
</tr>
<tr>
<td>Math score</td>
<td>-0.058</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.037)</td>
</tr>
</tbody>
</table>

**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 with just over 95000 students in total; 908 schools with at most two classes per grade with just over 60000 students in total; and 331 schools with one class per grade with just over 13000 students in total.

*, **, *** significant at the 10 percent, 5 percent, and 1 percent significance level respectively.
Table 7: Counterfactual Experiments of Lower Thresholds for Grade Retention

<table>
<thead>
<tr>
<th>Retention Rate</th>
<th>Uniform Distribution</th>
<th>Normal Distribution</th>
<th>Triangular Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grade Birth Cohort</td>
<td>Grade Birth Cohort</td>
<td>Grade Birth Cohort</td>
</tr>
<tr>
<td></td>
<td>Boys Girls</td>
<td>Boys Girls</td>
<td>Boys Girls</td>
</tr>
<tr>
<td>Boys Girls</td>
<td>0.09 0.41</td>
<td>0.09 0.42</td>
<td>0.06 0.47</td>
</tr>
<tr>
<td></td>
<td>(0.13) (0.12)</td>
<td>(0.13) (0.12)</td>
<td>(0.13) (0.13)</td>
</tr>
<tr>
<td>0.100 0.075</td>
<td>0.20 0.30</td>
<td>0.23 0.10</td>
<td>0.04 0.50</td>
</tr>
<tr>
<td></td>
<td>(0.12) (0.13)</td>
<td>(0.13) (0.13)</td>
<td>(0.13) (0.13)</td>
</tr>
<tr>
<td>0.075 0.054</td>
<td>0.21 0.31</td>
<td>0.26 0.09</td>
<td>0.06 0.47</td>
</tr>
<tr>
<td></td>
<td>(0.13) (0.13)</td>
<td>(0.13) (0.13)</td>
<td>(0.13) (0.13)</td>
</tr>
<tr>
<td>0.050 0.034</td>
<td>0.26 0.32</td>
<td>0.27 0.09</td>
<td>0.10 0.41</td>
</tr>
<tr>
<td></td>
<td>(0.13) (0.13)</td>
<td>(0.13) (0.14)</td>
<td>(0.13) (0.13)</td>
</tr>
<tr>
<td>0.025 0.016</td>
<td>0.29 0.27</td>
<td>0.29 0.10</td>
<td>0.17 0.31</td>
</tr>
<tr>
<td></td>
<td>(0.14) (0.13)</td>
<td>(0.14) (0.14)</td>
<td>(0.14) (0.14)</td>
</tr>
<tr>
<td>0.000 0.000</td>
<td>0.27 0.13</td>
<td>0.27 0.13</td>
<td>0.27 0.13</td>
</tr>
<tr>
<td></td>
<td>(0.15) (0.15)</td>
<td>(0.15) (0.15)</td>
<td>(0.14) (0.14)</td>
</tr>
</tbody>
</table>

Notes: The coefficients and standard errors in parentheses are averages across 100 simulations.
## Appendix Table 1: Descriptive Statistics

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

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