Working Paper No. 152

Girls' preferences for STEM and the effects of classroom gender composition: new evidence from a natural experiment

Damiano Pregaldini, Uschi Backes-Gellner and Gerald Eisenkopf

Universität Zürich
IBW – Institut für Betriebswirtschaftslehre

UNIVERSITÄT BERN

Working Paper No. 152

Girls' preferences for STEM and the effects of classroom gender composition: new evidence from a natural experiment

Damiano Pregaldini, Uschi Backes-Gellner and Gerald Eisenkopf

July 2020 (first version: June 2018)

This paper was previously circulated under the title "Students' Selection and Heterogeneous Effects of Classroom Gender Composition: Evidence from a Natural Experiment in Switzerland" (2018).

Published as: "Girls' preferences for STEM and the effects of classroom gender composition: New evidence from a natural experiment." *Journal of Economic Behavior & Organization*, 178(2020): 102-123. By Damiano Pregaldini, Uschi Backes-Gellner and Gerald Eisenkopf.

DOI: https://doi.org/10.1016/j.jebo.2020.07.018

Girls' preferences for STEM and the effects of classroom gender composition: new evidence from a natural experiment*

Damiano Pregaldini† Uschi Backes-Gellner‡ Gerald Eisenkopf§

July 16, 2020

Abstract

We analyze how preferences for STEM fields moderate the effect of classroom gender composition on the math grades of girls in high school. Using data from Switzerland, we compare students who have self-selected into a STEM specialization with students who have self-selected into a language specialization. Our identification exploits the random assignment of students to classrooms after they have chosen their specialization. In contrast to the average effects found in previous studies, we find a negative effect of the proportion of female peers in the classroom on math grades for girls who have self-selected into the STEM specialization and a positive effect for girls who have self-selected into a language specialization. These results are important for policies affecting the gender composition of classrooms.

Keywords: classroom gender composition, girls' preferences for STEM, students' selection

JEL Classification: I21, J16

*This study is partly funded by the Swiss State Secretariat for Education, Research, and Innovation (SERI) through its Leading House on the Economics of Education, Firm Behavior and Training Policies. Declarations of interest: none. We would like to thank Simone Balestra, Eric Bettinger, Thomas Dohmen, Tor Eriksson, Simon Janssen, Edward Lazear, Laura Rosendahl Huber, Guido Schwerdt, Carmit Segal, Niels Westergård-Nielsen, Conny Wunsch, Ulf Zölitz, seminar participants at the University of Zurich, participants at the annual Meeting of the Economics of Education Association 2018, the SASE Annual Conference 2018, and the 21st Colloquium on Personnel Economics for their valuable comments on this as well as earlier versions of this study. We would also like to thank Sara Brunner, Patricia Pálffy, and Marco Pereira for their excellent research assistance.

†University of Zurich, Graduate School of Business. E-mail: damiano.pregaldini@uzh.ch

‡University of Zurich, Department of Business Administration, backes-gellner@business.uzh.ch.

§University of Vechta, gerald.eisenkopf@uni-vechta.de.

1 Introduction

A large body of economic literature has analyzed the effect of classroom composition on educational production (Lazear, 2001; Anelli and Peri, 2017; Balestra et al., 2016; Patacchini et al., 2017; Buechel et al., 2018). The promise that educators can enhance educational production and improve school efficiency by altering the composition of classrooms has great appeal to many educational policy makers and researchers.

An important strand of this literature focuses on the effect of classroom gender composition on educational outcomes (Hoxby, 2000; Lavy and Schlosser, 2011; Whitmore, 2005). Although evidence from this literature is generally mixed (Oosterbeek and Van Ewijk, 2014; Hill, 2017; Schneeweis and Zweimüller, 2012; Black et al., 2013; Zölitz and Feld, 2017), studies focusing on students in primary and secondary education tend to consistently find that both girls and boys perform better, particularly in mathematics (hereafter, "math"), in classrooms with a higher proportion of girls. As Hoxby (2000) reports, this effect persists even after conditioning on peer ability, suggesting that the positive effect of having more girls in the classroom operates through changes not only in the quality of the peers but also in the classroom environment. Whitmore (2005) supports this result, emphasizing that "there is something about having girls in the class per se that improves outcomes" (p. 203). Lavy and Schlosser (2011) provide compelling evidence that this positive effect is mediated by an improved classroom environment. Specifically, having a higher proportion of female peers in the classroom reduces classroom disruption and violence, improves relationships among students and between students and the teacher, and reduces fatigue among teachers.

Another strand of the literature on gender effects in education studies the effect of single-sex vs. coeducational schools (Eisenkopf et al., 2015; Booth et al., 2014; Dustmann et al., 2018). These studies tend to find a positive average effect of single-sex schooling on students' grades, particularly in math. For example, Eisenkopf et al. (2015) study a pedagogical Swiss school in which students are randomly assigned to single-sex or coeducational classes, finding that assignment to an all-girls class has a positive effect on the math grades of female students—particularly those of high ability. Overall, these studies provide further evidence that the gender composition of classrooms matters for student outcomes.

Despite the growing literature on the average effects of classroom gender composition on different educational outcomes, little is known about how these effects vary across students with different preferences. One exception is Jackson (2012), who uses data from

Trinidad and Tobago to show, in a different context, that single-sex schooling improves achievement only among girls with a strong preference for single-sex classrooms. However, this is the only study addressing how the effect of classroom gender composition depends on individual student preferences. In particular, there is no evidence on how the effects of classroom gender composition depend on individual preferences for STEM (science, technology, engineering, and mathematics) fields, leading to heterogeneous results, for example, for girls with preferences for STEM vs. girls with preferences for other study fields such as languages. Given that gender segregation in schools has been advocated as a potential policy for reducing the gender gap in STEM fields, the lack of evidence on how the effect of classroom gender composition varies across students with different preferences for STEM is surprising. Our paper attempts to fill this gap.

In this paper, we analyze how classroom gender composition affects the grades of girls (and boys) with different preferences for STEM fields. To elicit students' preferences for STEM, we exploit students' self-selection into three specialization tracks (hereafter, "specializations") within one large Swiss gymnasium (a high school that grants direct access to university). Students have a choice among three common specializations: 1. STEM, 2. modern languages, and 3. ancient languages.

The self-selection of students into specializations (at grade 9, ages 14-15) reveals the students' true preference for one of the three specializations, as it is an important step affecting their future educational and career options (Wolter et al., 2014). As Lazear et al. (2012) show using laboratory experiments, accounting for individuals' self-selection into different environments is crucial when studying the effects of policy measures. Indeed, they show that the observed effects can vary dramatically (and, in some cases, can even reverse) across environments as a result of individuals' preferences and self-selection. Following a similar approach, we assume that the effect of classroom gender composition might vary across students with different preferences for STEM as expressed by their self-selection into a specialization.

To identify the causal effect of classroom gender composition on the grades of students with different preferences for STEM (i.e., in different specializations), we exploit a school policy that randomly assigns students to classrooms with naturally varying gender compositions after they have chosen their specialization (STEM, modern languages, or ancient languages). This policy generates exogenous variation in classroom gender composition in all three specializations. The focus on a specific school ensures that the students in the different specializations face the same environmental characteristics beyond their specific

specializations (e.g., facilities, regulations, and, in particular, teachers).

In contrast to the existing evidence, our results show that the effect of a higher proportion of girls in the classroom (hereafter, "proportion of girls") on math grades differs systematically across students with preferences for STEM or for languages: for students (girls and boys) who have self-selected into one of the two language specializations, we find, in line with previous studies, a positive effect of a higher proportion of female peers on math grades. In contrast, we find the opposite effect for girls who self-selected into the STEM specialization, i.e., those who revealed a preference for STEM over languages. These girls tend to have better math grades when the proportion of boys is higher.

Thus, our results reveal that the average positive effect of a higher proportion of female peers in the classroom found in other studies masks a high degree of heterogeneity across students with different preferences for STEM. Policy conclusions that draw from average treatment effects might therefore not be equally effective for all students. In our case, policy conclusions on gender classroom composition drawn from average effects would even point in the wrong direction for girls with a preference for STEM and might worsen the prospects for a STEM career at a very early stage.

Our analysis also shows that the effect on math grades is nonlinear for girls with a preference for languages. Specifically, for these students, the positive effect of increasing the proportion of female peers is particularly high in classrooms with few girls, but it decreases as the number of girls in the classroom increases. The existence of nonlinearities together with the finding that the effect is positive for some students but negative for others leaves scope for efficiency gains and allows us to derive suggestions for the optimal gender composition of classrooms within specializations.

To investigate possible explanations and mechanisms driving the effect heterogeneity, we survey the most recent cohort of high school students using a questionnaire. These data allow us to descriptively investigate students' characteristics and preferences and to derive possible explanations for why a higher proportion of female peers has a positive effect on girls in language specializations but a negative one on girls in the STEM specialization. Our results suggest that differences in the willingness to compete, in particular, might help explain the contrasting effects. Girls in the STEM specialization display higher levels of willingness to compete and might therefore benefit from a more competitive environment induced by an increase in the proportion of boys (i.e., a reduction in the proportion of female peers).

Our data show higher levels of competitiveness for students in the STEM specialization.

This finding is consistent with Buser et al. (2017) and is therefore not a peculiarity of our school. In contrast, girls with a preference for languages display, on average, lower levels of willingness to compete and might therefore benefit from a less competitive environment, i.e., an increase in the proportion of female peers.

Our results add to the literature analyzing the effect of classroom gender composition on students' achievement by showing that the sign and size of this effect differ not only by gender but also within gender by individual student characteristics and preferences for STEM. One implication of this finding is that no one-size-fits-all solution exists for the optimal gender composition of classrooms. Therefore, educational authorities should be cautious about considering classroom gender composition a universal solution for reducing the gender gap in math grades or for the underrepresentation of women in math-intensive STEM fields.

Our study also relates to the literature on the influence of female peers on the choice of a STEM track (e.g., Zölitz and Feld, 2017; Cools et al., 2019; Fischer, 2017; Mouganie and Wang, 2019). In line with our findings, these studies show that the effect of peers' gender differs across students with different characteristics. Our analysis thus adds to this literature by taking it one step further and analyzing how classroom gender composition affects student outcomes once student preferences for STEM are taken into account (by comparing students who have self-selected into a STEM specialization with those who have self-selected into a language specialization).

The remainder of the paper is structured as follows: Section 2 introduces the institutional background of the Swiss educational system to understand the process of self-selection into specializations and the exogenous variation in the classroom gender composition. Section 3 explains the empirical strategy. Section 4 describes the data and provides descriptive statistics. Section 5 presents the results and discusses the possible mechanisms of effect heterogeneity. Section 6 concludes the paper.

2 Institutional background

This section explains how classrooms are organized in the Swiss high school we analyze—a gymnasium that grants automatic access to Swiss universities. Our identification strategy relies on the students' self-selection into specializations combined with the school's random allocation of students to classes within specializations.

After compulsory schooling, which typically ends at age 15, adolescents in Switzer-

land enter the upper secondary level, where they choose among three types of education: vocational education and training (VET), specialized professional schools (i.e., schools providing general and professional education in specific occupational fields, e.g., healthcare), or gymnasium (i.e., academic high schools). On average, approximately 20% of Swiss adolescents choose the gymnasium. The gymnasium that we study is in the canton of Zurich[1] and has two levels: grades 7 and 8 ("Untergymnasium"—junior high school) and grades 9 through 12 ("Obergymnasium"—senior high school). Students typically enter high school in the junior level, i.e., grade 7 (ages 12-13), or in the senior level, i.e., grade 9 (ages 14-15). Before entering grade 9, each student has to choose one specialization among three options: STEM, modern languages, and ancient languages.[2] In grade 9, the school reassigns students to classes within the chosen specialization, whereupon students stay in the same class until graduation, i.e., from grades 9 through 12.

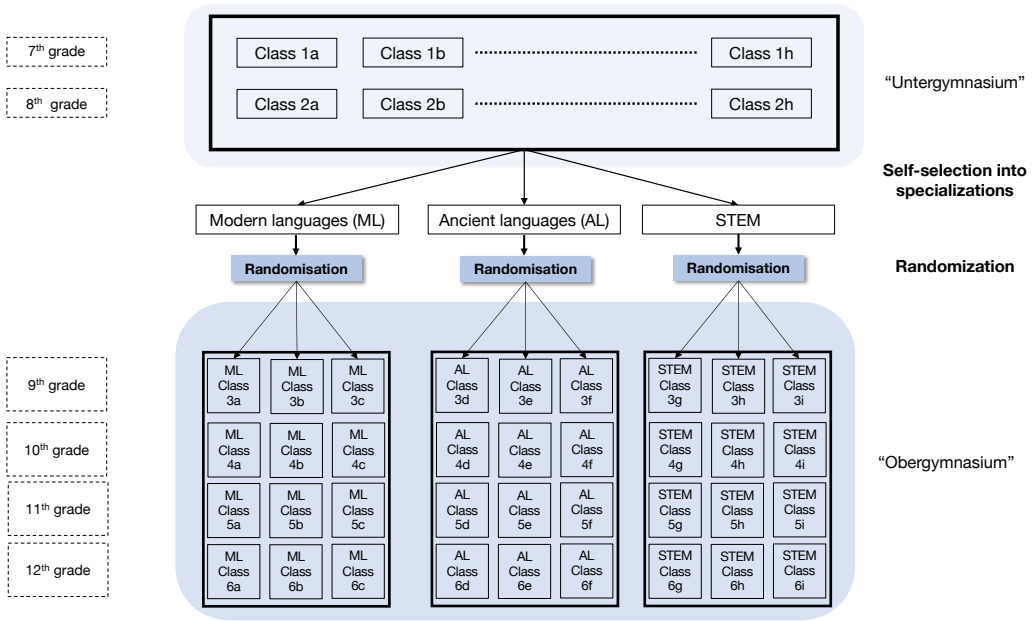
In each specialization, students take core courses (e.g., in math and German language studies) and specialized courses. In modern languages, students take specialized courses in English, Latin, Spanish, Italian, or Russian; in ancient languages, the classes are on Latin and ancient Greek; and in STEM, classes are in biology, chemistry, applied mathematics and physics.

After students have made their specialization choice, the school knows how many students will be studying in each specialization and determines the number of classrooms for each. Fig. 1 provides an illustrative example with three classrooms for each specialization (in reality, the number of classes within specializations varies between 2 and 5 in each cohort). In each specialization, the school administration randomly assigns students to classes (e.g., classes 3g, 3h, and 3i in Fig 1). The school administration carries out randomization to avoid both favoritism and parental politicking; it considers randomization to be the fairest mechanism to ensure a student classroom allocation that is independent of individual students' background and ability or the gender ratio. Therefore, the school administration applies this policy strictly and actively communicates and defends it against parents and students. The analysis of the data confirms that the process is random.

[1]Cantons are local government regions similar to U.S. states.

[2]The ancient languages specialization can be chosen only by students who started the high school in grade 7. Students who started in grade 9 can only choose between modern languages and STEM.

Figure 1: Structure of the classrooms in the school



Note: An illustrative example of the structure of the classrooms in the school. The actual number of classrooms might vary across cohorts and specializations.

3 Empirical strategy

We are interested in the possible causal effect of classroom gender composition on the grades of students with different preferences for STEM, i.e., in different specializations. A central problem in the estimation of such an effect is that the variation in classroom gender composition might be generated by selection (Hoxby, 2000). One concern is that students might choose classrooms with different gender compositions according to unobservable individual characteristics that also affect their grades. For example, girls with poor grades in math might feel more comfortable learning math in classrooms with a higher proportion of female peers. This type of selection, by unobservable student characteristics, would produce a spurious correlation between classroom gender composition and grades that might bias our estimate. Another concern is that parents might try to influence classroom assignment so that their children end up in a classroom with a gender composition that the parents consider more appropriate.

In our setting, we do not face this problem because the school randomly assigns stu-

dents to classes after they choose their specializations. Therefore, within specializations, students have no influence on the gender composition of the class to which they are assigned. In other words, this school policy ensures that individual preferences and ability are uncorrelated with the classroom gender composition within each specialization. Section 5.1 and Appendix A provide formal tests confirming this assumption.

As the high school has approximately 2,300 students and is quite large (in Swiss terms), we consider the randomization of classrooms a natural experiment providing a source of random variation in classroom gender composition that is exogenous to student ability within specializations. We exploit this variation to study how a higher proportion of female peers affects the grades of students in different specializations. Appendix A provides a formal test and strongly supports the argument that variation in classroom gender composition is exogenous to student ability.

Moreover, by focusing on one school, we ensure that students in different specializations face the same learning environment and educational inputs. Indeed, school building, classrooms, class schedules, and school curriculum (except for specialized subjects) are common to all students, irrespective of their specialization. Most importantly, students face the same pool of teachers across all three specializations, as teachers are not recruited to teach exclusively in a specific specialization. The data show that teachers often teach classes in different specializations. Moreover, all teachers must fulfill the same requirements to become high school teachers, irrespective of the specialization track they teach in. This setting allows us to compare students who are exposed to the same environment, institutional setting, and resources. Therefore, differences in these factors are unlikely to determine the heterogeneity across specializations.

Our setting allows us to exploit the variation in the proportion of female peers both across cohorts—as in Hoxby (2000), Lavy and Schlosser (2011), and Black et al. (2013)—and across classrooms (as the random assignment prevents students from sorting into classrooms within the specialization). Moreover, as approximately 90% of the teachers in the high school taught more than one class during our observation period, we are able to include teacher fixed effects to absorb time-invariant heterogeneity at the teacher level and capture systematic differences in grading among teachers. The linear specification for the grade of student i in classroom c, specialization k, with the subject-specific teacher s, and in semester t is:

$$score_{ickst} = \alpha + \beta ratio^{-i}_{ckst} + \gamma female_{ickst} + \delta clsssize_{ckst} + \theta_k + \eta_s + \lambda_t + \epsilon_{ickst} \quad (1)$$

where $ratio^{-i}_{ckst}$ is the leave-one-out proportion of female peers in student i's classroom c, specialization k, with teacher s, and semester t; $female_{ickst}$ is a gender dummy; and $clsssize_{ckst}$ is the size of student's i class. Because randomization takes place within specializations, we include specialization fixed effects θ_k. η_s and λ_t are teacher and semester fixed effects. Because our data allow us to identify teachers in math and German language studies separately, η_s is a subject-specific teacher fixed effect. Finally, we cluster standard errors at the classroom level c.

The parameter β identifies the causal effect of having a higher proportion of female peers in the classroom on grades. In our main specification, we drop the gender dummy and the specialization fixed effects, and we separately estimate Eq. 1 in subsamples of girls and boys in different specializations.

4 Data

For our empirical analysis, we use school internal records including all students (i.e., grades 7 to 12) who were enrolled in the high school from September 2002 through June 2012. These records provide individual-level information on semester grades, gender, specialization, and the gender composition of each classroom. Moreover, the data include classroom and teacher identifiers, which allow us to identify students in the same classroom and students taught by the same teacher over time.

These administrative school records are particularly well suited to our investigation for three reasons. First, in contrast to some previous studies (e.g., Hoxby, 2000), the classroom identifier allows us to define gender composition at the classroom level. As students stay in the classroom throughout the day, attending the vast majority of lessons together with the same peers, the classroom represents an ideal setting for studying gender effects. Second, we are able to observe student selection into one of the three specializations, a selection that reveals their educational preferences. Third, the high school district encompasses not only parts of the city of Zurich but also nearby towns and villages. Therefore, the large number of students and the mix of individuals with rural and urban backgrounds increases the representativeness of our sample for the population of Swiss high school students.

Table 1 reports the total number of students, observations, and classrooms in each cohort. In each cohort, we observe between 127 and 209 students in six to ten classrooms. Because we observe students in the high school from September 2002 through June 2012, some cohorts have more observations than others. For example, for the class of 2003, we observe only the 12th grade, while for the class of 2015, we observe only the 9th grade. Because some students may repeat terms, leave the high school due to unsatisfactory academic performance, move away, or enter the high school after grade 9, our dataset takes the form of an unbalanced panel.[3]

Table 1: Structure of the data

Class of	Students	Observations	Classrooms
2003	127	200	6
2004	143	492	8
2005	172	870	8
2006	174	1100	8
2007	180	1270	8
2008	172	1190	8
2009	170	1214	8
2010	183	1335	8
2011	172	1265	8
2012	186	1192	8
2013	182	990	8
2014	209	796	10
2015	209	414	10
Total	2279	12328	106

Note: Number of students, observations, and classrooms in each cohort. "Class of" indicates the expected year of graduation.

As randomization takes place in the transition from 8th to 9th grade, we restrict the sample to students in the 9th through 12th grades for our analysis. Because we observe these students over several semesters between the 9th and 12th grades, each observation in our estimation sample corresponds to one student in one semester. After we account for missing values, the sample contains 12,328 observations over 2,279 students. Because

[3]Seventy-five percent of the observations in our estimation sample are from students who remain in our dataset until graduation.

31 students in the sample (1.4 percent of the sample) switch specialization during or after 9th grade, we observe a slightly higher number of student-specialization pairs (2,310) than number of students: 844 of them (36 percent) in modern languages, 592 (26 percent) in ancient languages, and 874 (38 percent) in STEM.

4.1 Classroom gender composition: Main explanatory variable

The explanatory variable of interest is classroom gender composition. To better separate the effect of the student's own gender from the effect of the peers' gender, we operationalize classroom gender composition as the classroom leave-one-out mean of the gender variable, i.e., the proportion of female peers in the classroom:

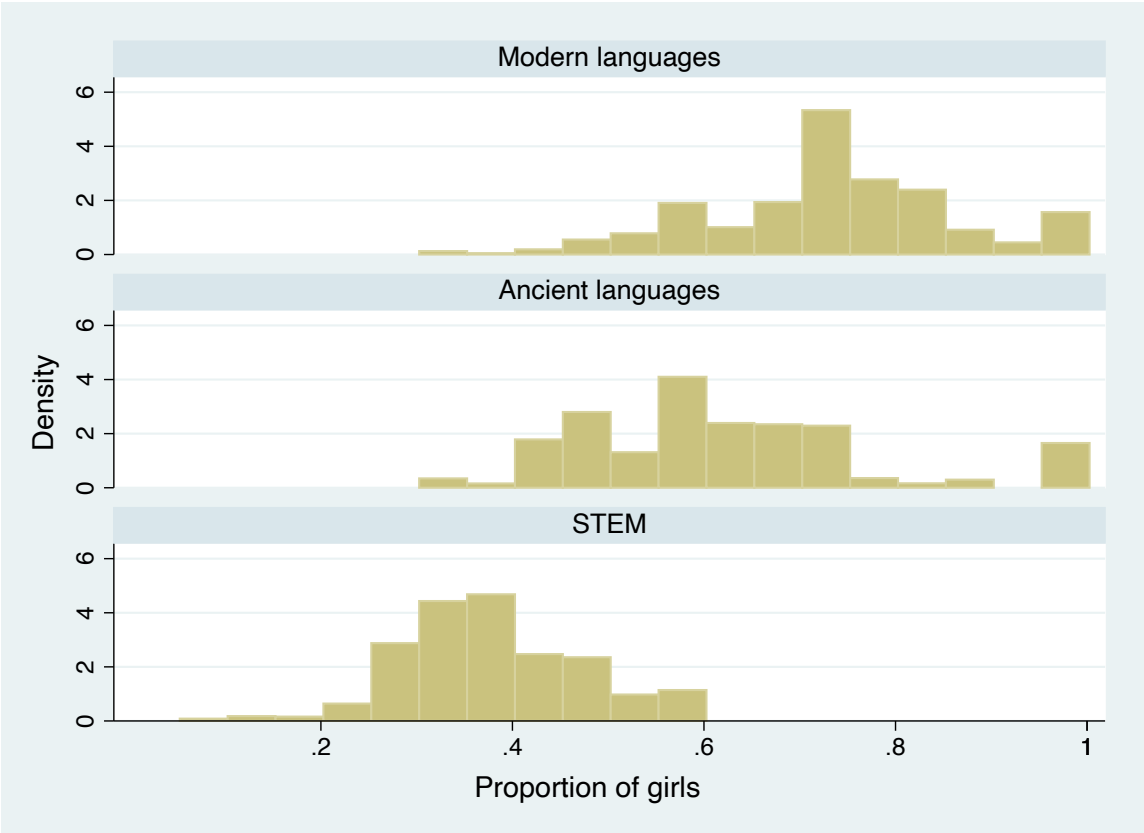
$$ratio_{ckst}^{-i} = \frac{1}{N_c - 1} \sum_{j \neq i}^{N_c} female_{jckst} \tag{2}$$

where c indicates the classroom, k the specialization, s the subject-specific teacher, and t the semester. The subscript j indicates all students in classroom c who are not i (i.e., all of student i's classmates in classroom c). Finally, N_c is the number of students in classroom c.

The gender composition of classrooms varies strongly within and across specializations. As Fig. 2 illustrates, the average proportion of girls is particularly high in modern (73.6 percent) and ancient languages (62.6 percent), ranging from 31 to 100 percent. In contrast, the average proportion of girls in STEM classrooms is low (38.1 percent), ranging from 5 to 59 percent. Fig. 2 also shows considerable variation in classroom gender composition within specializations. Given that we exploit the variation within specializations, this variation is important.

Although the average proportion of girls in the classroom is different between the STEM and language specializations, Figure 2 shows large overlaps in the gender distribution across specializations. These overlaps imply that selection into STEM or languages does not automatically lead to a high or low proportion of girls in the classroom. In fact, the choice of a specialization does not guarantee a specific classroom gender composition.

Figure 2: Proportion of girls in different specializations



Note: Every observation is the proportion of girls in a classroom of a given cohort and semester. Modern languages: Mean=73.6%, SD=13.2. Ancient languages: Mean=62.6%, SD=15.6. STEM: Mean=38.1%, SD=9.5.

4.2 Grades in math and German language studies: Outcome variables

Our outcome variables of interest are student grades in math and German language studies (the students' language of instruction). School achievement, especially in math, is an important predictor of future earnings (e.g., Bertrand et al., 2010). Following (Hoxby, 2000) and (Eisenkopf et al., 2015), we focus our analysis on student achievement in math and in the native language (German), as expressed by grades in these subjects. Given that Switzerland has no standardized test at the end of each grade, we follow Eisenkopf et al. (2015) and use grades from student report cards. In Switzerland, grading ranges from 1 to 6, where 1 is the lowest grade, 4 is the passing mark, and 6 is the highest grade. Fig. 3

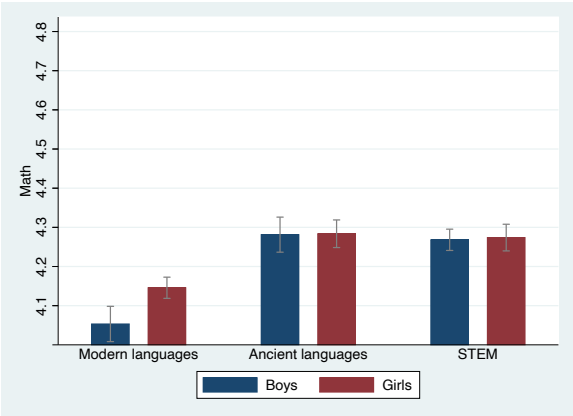


Figure 3: Test scores in math (higher scores are better scores, with 6 being the maximum score and 4 being the lowest passing grade)

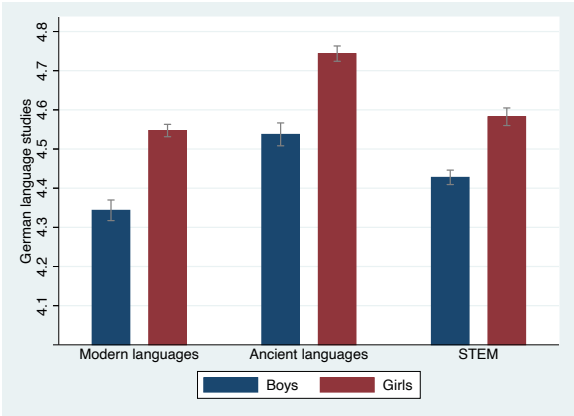


Figure 4: Test scores in German language studies (higher scores are better scores, with 6 being the maximum score and 4 the lowest passing grade)

and Fig. 4 report grades in math and German language studies in grades 9-12.

Consistent with previous studies (e.g., Guiso et al., 2008), we find that girls outperform boys in the languages (in all three specializations). In math, girls in modern languages outperform boys on average. However, in ancient languages and STEM, girls' average grades do not systematically differ from boys'.

5 Results

In this section, we first test the validity of our identification strategy and then present the results of the estimation of Eq. 1 for the subsamples of boys and girls in different specializations.

5.1 Validity of the identification strategy

Our identification strategy relies on the random assignment of students to classrooms within specializations in the transition from 8th to 9th grade. However, if students were to sort within specializations into classrooms with different gender compositions according to either ability or other unobservable characteristics related to grades, the coefficient β in Eq. 1 would no longer capture the causal effect of having a larger proportion of female peers in the classroom on grades.

To assess the validity of our identification strategy, we test for each cohort and specialization (hereafter, "specialization-by-cohort cell") whether randomly formed 9th grade classrooms differ in terms of predetermined student characteristics. We focus on three student characteristics that we can observe in the school register data: gender and the last available grades in math and German language studies prior to randomization (i.e., in 8th grade), as proxies for ability.

To implement this test, in each 9th-grade specialization-by-cohort cell, we regress the student characteristic on a battery of classroom fixed effects. If assignment is random, we expect that classrooms in the same specialization-by-cohort cell will not systematically differ in terms of predetermined student characteristics (i.e., the classroom fixed effects will be jointly nonsignificant). For gender, this approach is analogous to testing whether classrooms gender composition—our treatment of interest—systematically differs across classrooms in the same specialization-by-cohort cell.

Table 2 reports the number of specialization-by-cohort cells for which we reject the H_0 of no difference in average grades or gender composition across classrooms at the three conventional significance levels (1, 5 and 10 percent). The column "Expected" reports the number of cells in which we would expect to detect differences across classrooms if assignment was completely random. To obtain this number, we multiply the significance level by the total number of cells for which we run the test (30 for gender and 27 for grades).

For grades in math and German language studies prior to randomization, the number of cells for which the test detects differences across classrooms closely matches the expected number of cells. This result supports the assumption that students are assigned to classrooms independently of their ability. For the students' gender, we detect differences in a number of cells that is higher than expected. At the 1 percent level, we detect differences in 5 cohort-by-specialization cells (out of 30), while we would expect to detect differences in only 0.3 cells, on average.

A closer inspection of the classrooms in the specialization-by-cohort cells in which we detect significant differences shows that these cells are either in modern languages or ancient languages. For STEM, we do not detect any difference in the gender composition of classrooms in the same cell. Moreover, four of the five specialization-by-cohort cells in which we detect differences contain either single-sex (all female) classrooms or mixed-specialization classrooms (i.e., classrooms with some students specializing in ancient languages and others specializing in modern languages).

Table 2: Test for random assignment of students to classrooms

	Number of specialization-by-cohort cells					
	Math	German language studies		Gender		
Significance level	Actual	Actual	Expected	Actual	W/o single-sex and mixed-specialization classrooms	Expected
1%	1	0	0.27	5	0	0.3
5%	2	1	1.35	5	0	1.5
10%	3	2	2.7	6	0	3

Note: The columns "Actual" report the number of specialization-by-cohort cells for which the test detects differences across classrooms with respect to grades or the gender composition. The columns "Expected" report the number of specialization-by-cohort cells for which the test is expected to detect differences across classrooms under perfect randomization. This number is obtained by multiplying the significance level by the total number of specialization-by-cohort cells. The total number of specialization-by-cohort cells is 30 for gender (3 specializations and 10 cohorts) and 27 for grades (3 specializations and 9 cohorts; we exclude the first cohort because we do not observe the grades of these students prior to randomization). In the column "W/o single-sex and mixed-specialization classrooms", we run the test on a subsample of students excluding single-sex and mixed-specialization (modern *and* ancient languages) classrooms.

The presence of single-sex classrooms (3 out of 84 9th-grade classrooms are single-sex, one in modern languages and two in ancient languages) and mixed-specialization classrooms (roughly 9 percent the estimation sample) might result in classrooms with systematically different gender compositions within the same specialization-by-cohort cell. In the column "W/o single-sex and mixed-specialization classrooms", we exclude all three single-sex classrooms and three mixed-specialization classrooms in specialization-by-cohort cells for which we detect significant differences in the classroom gender composition. Once we exclude these classrooms, we detect differences in zero cells, indicating that students are assigned to classrooms independently of their gender. In Appendix C, we test the robustness of our results with respect to the exclusion of single-sex and mixed-specialization classrooms. The results show that excluding these classrooms has little effect on our estimates.

Together, these tests substantiate the argument that students are randomly assigned to

classrooms within specializations, supporting our key identifying assumption that variation in classroom gender composition is as good as random within specializations. In appendix A, we provide further evidence supporting this assumption.

5.2 Grades in math and German language studies

Table 3 reports the coefficients of the regression of individual grades in math (Panel A) and German language studies (Panel B).

Table 3: Boys and girls in different specializations

	Modern Languages		Ancient Languages		STEM	
	boys	girls	boys	girls	boys	girls
Panel A: math						
Proportion of female peers	0.176***	0.101***	−0.003	0.090***	−0.001	−0.120***
	(0.033)	(0.036)	(0.054)	(0.029)	(0.027)	(0.035)
Adj.-R2	0.096	0.056	0.016	0.068	0.015	0.064
Panel B: German language studies						
Proportion of female peers	−0.064	0.040	−0.026	0.050	0.059	−0.110**
	(0.051)	(0.035)	(0.090)	(0.035)	(0.037)	(0.051)
Adj.-R2	0.120	0.046	0.082	0.048	0.044	0.096
Semester FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Teacher FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Controls	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Obs.	1163	3252	1283	2145	2780	1705
Students	233	611	226	366	555	319

Note: Each coefficient represents a separate regression in the subsamples of boys and girls in different specializations. Standard errors are in parentheses. Standard errors are clustered at the classroom level. Each observation is a student in a semester. Grades are standardized by specialization and gender. The proportion of female peers (leave-one-out mean) is standardized by specialization: the estimates report the effect (in standard deviations) of a one standard deviation increase in the proportion of female peers. Controls: class size. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The coefficients in Panel A show that for girls in the STEM specialization, the effect of the proportion of female peers is negative: one standard deviation (10 percentage points, i.e., roughly 2-3 girls in an average classroom of 22 students) increase in the proportion of female peers decreases math grades by 0.12 standard deviations. In contrast, the effect is positive for almost all students in language specializations (except for boys in ancient languages): one standard deviation increase in the proportion of female peers increases math grades by 0.18 standard deviations for boys in modern languages, 0.10 for girls in modern languages, and 0.09 for girls in ancient languages.

Thus, the effect of having more girls in the classroom is mainly positive for students with a preference for languages and negative for girls with a preference for STEM fields. These results show that the effects of a higher proportion of female peers on math grades systematically differ across students who have self-selected into different specializations.[4]

To put our estimates in perspective, we compare them with estimates from other studies on gender effects in schools. Hoxby (2000) uses data on 3rd to 6th graders in Texas and finds that a 10-percentage-point increase in the proportion of female peers in the same cohort increases math grades by 1 to 3 percent of a standard deviation (own calculations using the information reported in Hoxby, 2000, Table 3, p. 33 and Table 5, p. 35). Lavy and Schlosser (2011) use data on Israeli high school students and find that a 10-percentage-point increase in the proportion of female peers increases students' average scores by 2 to 2.5 percent of a standard deviation. Depending on the specialization, our estimate in the full sample (Table 10, Column 3) implies that a 10-percentage-point increase in the proportion of female peers improves math grades by two to five percent of a standard deviation. This estimate is in line with the results in Hoxby (2000) and Lavy and Schlosser (2011). However, neither Hoxby (2000) nor Lavy and Schlosser (2011) find gender-specific responses of grades to classroom gender composition.

In contrast, Brenøe and Zölitz (forthcoming) and Zölitz and Feld (2017) find gender-specific responses to peer gender composition. Brenøe and Zölitz (forthcoming) find that a 10-percentage-point increase in the proportion of female peers increases the GPA of male students by 1.3 percent of a standard deviation but has no effect on the GPA of female students. Zölitz and Feld (2017) find that female students benefit from female peers

[4] Appendix B tests whether the coefficients in Table 3 are significantly different from one another. Table 10 reports the estimation results in the full sample, where the proportion of female peers is interacted with the specialization. Table 10 shows that the effect is positive in languages but not in STEM. Table 11 tests whether the coefficients reported in panel A of Table 3 are significantly different from one another.

only in nonmathematical courses, while male students benefit from female peers only in mathematical courses. Specifically, a 10-percentage-point increase in the proportion of female peers increases the GPA of female students by 1.2 percent of a standard deviation in nonmathematical subjects and the GPA of male students by 1.3 percent of a standard deviation in mathematical subjects.

Consistent with these findings, our results show gender-specific responses to classroom gender composition in some specializations. For example, the proportion of female peers has a negative effect on the math grades of girls in STEM but does not affect the math grades of boys in the same specialization. However, our estimated coefficients for the subsamples of boys and girls in different specializations (Table 3) tend to be larger than those reported in the literature. These coefficients indicate that the effects of peer gender composition are heterogeneous not only between genders but also (and potentially more importantly) across students in different specializations, that is, students with different preferences and characteristics. More generally, the heterogeneity in how students respond to the gender composition of their peer group might help explain why studies reach different conclusions about gender peer effects (e.g., Anelli and Peri, 2017).

In addition, to investigate potential nonlinearities in the relationship between the proportion of female peers and math grades, we run a fractional polynomial regression in the spirit of Royston and Altman (1994). The results (cf. Appendix D) suggest that the effect is nonlinear for boys and girls in modern languages and for girls in ancient languages. Specifically, for boys in modern languages, the effect appears to increase with the number of girls in the classroom. In contrast, for girls in both the modern and ancient language specializations, an increase in the proportion of female peers produces larger positive effects when there are few girls in the classroom. However, the positive effect decreases as the number of girls in the classroom increases. These findings indicate that for girls in language specializations, classrooms with few girls benefit the most from an increase in the proportion of female peers, while the opposite is true for boys in modern languages.

The results for the grades in German language studies (panel B, Table 3) again show a negative effect of the proportion of female peers for girls in the STEM specialization. However, we do not find any significant effect on the grades of girls in the other specializations.

For more insights into the heterogeneous effects of classroom gender composition, the next section discusses potential mechanisms. The discussion will focus on math grades, as Table 3 shows little evidence of a significant relationship between classroom gender composition and grades in German language studies.

5.3 Mechanisms

The most counterintuitive finding is that the effect of the proportion of female peers on math grades is positive for the girls who self-selected into the language specializations (both modern and ancient languages) but negative for the girls who self-selected into the STEM specialization. Thus, the question arises as to what mechanisms could drive these contrasting effects. This section therefore compares girls in the languages and STEM specializations and discusses how the self-selection of girls into the respective specialization might drive the contrasting effects of classroom gender composition.

To analyze differences across girls who self-selected into different specializations, we use survey data that we were able to collect from the most recent cohort of 12th graders in our data set (graduation year 2014) in all three specializations.[5] We focus on the willingness to compete, as an increasing number of studies have shown the importance of competitiveness in explaining different types of gender gaps (e.g., Niederle and Vesterlund, 2007), including the gender gap in STEM (e.g., Buser et al., 2014, 2017). In addition, we provide a brief overview of some potential alternative explanations.

5.3.1 Willingness to compete

One characteristic that differentiates girls in the STEM specialization from girls in the language specializations is their willingness to compete.[6] As (Buser et al., 2014) show, students with a higher willingness to compete tend to select into more math-intensive and prestigious specializations in high school. This finding is confirmed by (Buser et al., 2017), who find, in a setting similar to ours, that students with a higher willingness to compete are more likely to select into the STEM specialization.

One result of this self-selection process is that students in the STEM specialization have, on average, a higher willingness to compete than students in the language specializations.

[5]The survey provides information on the students' interest in and attitude toward math and German language studies, the classroom environment, and personal and family socioeconomic background. A total of 174 students in all three specializations completed the questionnaire in December 2014: 101 in the languages (modern and ancient languages together) and 73 in STEM. Appendix E provides more details on the survey

[6]Some recent research suggests that gender differences in other traits, such as risk preferences and overconfidence, largely explain gender differences in competitiveness (e.g., van Veldhuizen, 2017). We focus on differences in competitive behavior as expressed by the answer to the following question: "Are you generally a person who enjoys competing with others or one who tries to avoid situations where you have to compete with others? "

These differences in the willingness to compete might, in turn, affect how students in different specializations respond to the gender composition of the classroom.

Given our empirical results, differences in the willingness to compete across specializations would imply that girls in the STEM specialization with a higher willingness to compete benefit from a more competitive classroom environment (i.e., a higher proportion of boys in the classroom because boys are, on average, more competitive than girls—as shown by (Gneezy and Rustichini, 2004; Gneezy et al., 2003)), while girls in the language specializations with a lower willingness to compete benefit from a less competitive environment (i.e., a higher proportion of girls because girls are, on average, less competitive than boys). Therefore, differences in the willingness to compete might act as an underlying mechanism for the different effects of classroom gender composition on the math grades of girls across specializations.

Although the internal school records do not allow us to directly test this mechanism, we can descriptively analyze with our survey data whether girls in the STEM specialization are more willing to compete than girls in the language specializations. Table 4 reports the average willingness to compete in the full sample and in the subsamples of boys and girls in the STEM and language specializations. We find that students in the STEM specialization are more willing to compete than students in the language specializations. This result is consistent with Buser et al. (2014) and Buser et al. (2017) for Swiss high school students. Moreover, we observe that boys are more willing to compete than girls. This finding is consistent with the idea that an increase in the proportion of boys is associated with higher levels of competitiveness in the classroom, whereas an increase in the proportion of female peers is associated with lower levels of competitiveness in the classroom.

Our data further show that girls in the STEM specialization are more willing to compete, on average, than girls in the language specializations. This difference is consistent with our argument that more competitive girls in the STEM specialization benefit from a more competitive environment triggered by a higher proportion of boys in the classroom, while the opposite is true for girls in the language specializations. However, this mechanism should be further explored in future research, as the difference is not significant (possibly because of the low number of observations).

Table 4: Differences in the willingness to compete

	Willingness to compete		
	Languages	STEM	Δ
Full sample	4.47	5.80	-1.33***
			(0.43)
N	101	73	
Girls	4.10	4.75	-0.65
			(0.63)
N	73	26	
Boys	5.45	6.38	-0.93
			(0.63)
N	28	47	

Average willingness to compete of girls and boys in different specializations. Students in the modern and ancient language specializations are pooled in the group "Languages". Scale: 1="I do not like to engage in competition at all" to 10="I really like engaging in competition". Standard errors for the difference in means in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.3.2 Further mechanisms

Girls in different specializations differ not only in their willingness to compete but also in other characteristics. Therefore, we provide a brief overview of alternative mechanisms for which we can find some evidence in our data. Lavy and Schlosser (2011) find that lower levels of classroom disruption and violence, improved relationships among students and between students and teachers, and lower teacher fatigue mediate the positive effect of a higher proportion of girls on grades.

Given our results, we would expect that this moderating effect operates differently in the STEM specialization classrooms compared to the language specialization classrooms. For example, students with little interest in the subject are possibly more likely to be both disruptive and easily distracted by disruptive classmates. In contrast, highly motivated students with a strong interest in the subject are likely to be more focused, less disruptive, and less affected by disruptive behavior in the classroom Bru (2006).

Indeed, our data reveal that girls who self-selected into the STEM specialization have a stronger interest in math, are more confident about their ability in math, and rate their math lessons as less noisy than girls who self-selected into the language specializations. Thus, girls in the STEM specialization might be less disrupted when there are more boys in the classroom than girls in the language specializations. These differences might therefore contribute to explaining the heterogeneity in the effect of classroom gender composition on math grades. Future research might provide more insights into the mechanisms driving the contrasting effects of classroom gender composition.

Differences in socioeconomic characteristics might also drive the effect heterogeneity across specializations. For example, Brenøe and Zölitz (forthcoming) find that parental education moderates the effect of peers on female students. Appendix E investigates the differences in socioeconomic characteristics across students in different specializations. While students in different specializations do not appear to differ in terms of parental education, students in STEM, on average, have more PCs, live in larger houses, and have fewer cars than students in modern languages.

6 Robustness checks

6.1 Nonlinear effects

This section addresses the concern that the effect heterogeneity is driven by a U-shaped relationship between math grades and the proportion of female peers rather than by differences across students who self-selected into different specializations. For example, an increase in the proportion of female peers might have a positive effect on students' grades if the number of girls in the classroom is already high (as in modern and ancient languages) but a negative effect if the number of girls is low (as in STEM).

The analysis of nonlinearities in appendix D provides the first empirical evidence that speaks against this alternative explanation. Indeed, for girls in modern and ancient languages (figures 5b and 5d), we observe the opposite relationship: the effect of the proportion of female peers is high when there are few girls in the classroom, but the effect decreases as the number of girls increases.

To further substantiate the argument that our results are not driven by differences in the distribution of girls across specializations, we exclude from our estimation sample all classrooms that we cannot compare across specializations because the proportion of female

peers is either smaller or larger than all the proportions in the other specializations. We therefore run the base model on two restricted samples of students. In the first sample, we include students in classrooms with a proportion of female peers of between 30 and 60 percent (cf. Fig 2). In the second sample, we include students in classrooms with a proportion of female peers of between 40 and 60 percent, thereby further restricting the estimation sample, particularly for the students in STEM. Table 5 reports the results of the regressions in the two restricted samples.

When we restrict the sample to classrooms with a proportion of girls between 30 and 60 percent (Panel A), the sign of the estimated coefficients is similar to those obtained in the full sample (Table 3). However, because we restrict the variance of the explanatory variable, the estimated coefficients tend to be larger. When we further restrict the sample to students in classrooms with a proportion of girls between 40 and 60 percent, the estimate coefficients do not qualitatively change. However, the coefficient for girls in the modern languages becomes marginally nonsignificant (p-value: 0.12). Taken together, these results indicate that differences in the proportion of girls are unlikely to drive the effect heterogeneity across specializations.

Table 5: Restricted sample

	Modern Languages		Ancient Languages		STEM	
	boys	girls	boys	girls	boys	girls
Panel A: Proportion of girls between 30 and 60 percent						
Proportion of female peers	0.478***	0.535***	0.060	0.375**	0.008	−0.130***
	(0.080)	(0.164)	(0.076)	(0.137)	(0.025)	(0.045)
Obs.	350	389	818	835	2220	1521
Students	90	104	180	207	436	286
Adj.-R2	0.102	0.124	0.025	0.113	0.025	0.068
Panel B: Proportion of girls between 40 and 60 percent						
Proportion of female peers	0.593*	0.326	0.086	0.282**	0.002	−0.193*
	(0.268)	(0.189)	(0.093)	(0.101)	(0.067)	(0.096)
Obs.	323	377	775	812	911	826
Students	90	104	180	207	240	200
Adj.-R2	0.095	0.082	0.024	0.091	0.011	0.040
Semester FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Teacher FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Controls	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*

Note: Each coefficient represents a separate regression for the subsamples of boys and girls in different specializations. Standard errors are in parentheses. Standard errors are clustered at the classroom level. Each observation is a student in a semester. Grades are standardized by specialization and gender. The proportion of female peers (leave-one-out mean) is standardized by specialization: the estimates report the effect (in standard deviations) of a one standard deviation increase in the proportion of female peers. Controls: class size. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6.2 Cohort variation in the proportion of girls

Starting with Hoxby (2000), a number of studies have used cross-cohort variation in the proportion of girls to identify the effect of having more females in a peer group (e.g., Lavy and Schlosser, 2011; Black et al., 2013). While our main strategy exploits variation in peer gender composition across both cohorts and randomly formed classrooms, this alternative approach exploits only the part of variation in peer gender composition that comes from

variation across cohorts. This approach builds on the assumption that cohort-to-cohort variation in the proportion of girls is generated by variation in the gender of births and it is therefore unrelated to unobserved determinants of student achievement.

To test the robustness of our result, we follow Hoxby (2000, p.8) and estimate the effect of the proportion of girls using variation within specializations and across cohorts. Specifically, we estimate the following model in first differences:

$$\Delta\overline{grade}_{gkt,c} = \beta\Delta ratio_{kt,c} + \Delta\epsilon_{gkt,c} \tag{3}$$

The subscript g refers to the gender (boys, girls); k to the specialization (modern languages, ancient languages, and STEM); and t to the semester (from semester 5, i.e., the first semester after randomization to semester 12). $\Delta\overline{grade}_{gkt,c} = \overline{grade}_{gkt,c} - \overline{grade}_{gkt,c-1}$ is the difference in the average grades of the group identified by the subscripts g, k, and t in two adjacent cohorts. $\Delta ratio_{kt,c} = ratio_{kt,c} - ratio_{kt,c-1}$ is the difference in the proportion of girls in specialization k and semester t between two adjacent cohorts. Finally, $\Delta\epsilon_{gkt,c}$ is the idiosyncratic error term. Under the assumption that $\Delta ratio_{kt,c}$—the cohort-to-cohort change in the semester- and specialization-specific proportion of girls—is unrelated to students' characteristics, β uncovers the causal effect of having a more female peer group on grades.

Table 6 reports the estimates of the coefficient in Eq. 3 for the grades in math and German language studies of subgroups of boys and girls in different specializations. Because the observations are cohort-by-semester-by-specialization group averages and the groups vary in size, we weight each observation by the size of the corresponding group. As in our main specification, we standardize the proportion of girls in the classroom by specialization.

The coefficients represent the effect of a one standard deviation increase in the first differences in the proportion of girls (i.e., the difference in the proportion of girls within specializations and semesters between two adjacent cohorts) on the first differences in grades (in absolute points). One standard deviation in the first differences in the proportion of girls is 10 percent in modern languages, 16 percent in ancient languages, and 12 percent in STEM. For comparison with our main model (Table 3), we state in square brackets the estimated coefficients in terms of the standard deviations of the grades of the respective group.

This alternative strategy provides estimates that are qualitatively similar to those of our main specification in Table 3. However, they appear to be larger and more significant

Table 6: Cohort variation

	Modern Languages		Ancient Languages		STEM	
	boys	girls	boys	girls	boys	girls
Panel A: math						
Δ ratio	0.110***	0.046	0.070**	0.156***	0.095***	−0.052
	(0.037)	(0.030)	(0.032)	(0.027)	(0.022)	(0.032)
	[0.14]	[0.06]	[0.09]	[0.19]	[0.13]	[−0.07]
Adj.-R2	0.099	0.019	0.051	0.320	0.195	0.023
Panel B: German language studies						
Δ ratio	0.014	0.030**	0.168***	0.039**	0.066***	−0.082***
	(0.022)	(0.014)	(0.026)	(0.018)	(0.015)	(0.027)
	[0.03]	[0.06]	[0.32]	[0.08]	[0.13]	[−0.17]
Adj.-R2	0.000	0.051	0.367	0.050	0.211	0.104
Obs.	71	71	71	71	71	71

Note: Each coefficient represents a separate regression for the subsamples of boys and girls in different specializations. Standard errors are in parentheses. The proportion of girls is standardized by specialization: the estimates report the effect (in points) of a one standard deviation increase in the proportion of female peers. Each observation is weighted by the size of the respective specialization-by-semester-by-cohort group. In square brackets, we state the estimated coefficients in terms of the standard deviations of the grades. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in most of the subgroups. For math, the estimated coefficients are positive in almost all subgroups except for girls in STEM. For this subgroup, the effect is again negative albeit marginally nonsignificant (p-value: 0.11). For girls in modern languages, the effect is again positive but imprecisely estimated (p-value: 0.13). For German language studies, we find positive effects in all subgroups except for girls in STEM and boys in modern languages. For girls in STEM, the estimated coefficient is negative and significant. Overall, the estimates from our main specification appear to be more conservative than those from this alternative strategy.

6.3 Proportion of newcomer girls

To further test the robustness of our results, we use an alternative source of variation in classroom gender composition. Specifically, we focus on the variation in classroom gender

composition generated by newcomer girls in 9th grade, that is, girls coming from other schools.

Because newcomer students can, in principle, specialize only in modern languages and STEM, we run this test on these two specializations only. In so doing, we create two new variables measuring the proportion of newcomer girls in the classroom. The first variable measures the proportion of girls in the classroom using only those girls who arrived in 9th grade and were randomly assigned to a classroom. Across all cohorts, we observe 205 girls in modern languages and 91 in STEM who arrived in 9th grade only.

The second variable measures the proportion of girls in the classroom using all girls who arrived in or after 9th grade. That is, we exploit the variation in classroom gender composition generated by girls who were assigned to a classroom in or after 9th grade. Across all cohorts, we observe 361 girls in modern languages and 158 in STEM who arrived in 9th grade or later. Table 7 reports the results of the two regressions using the alternative definitions of classroom gender composition. For brevity, we present the results only for math grades (we report the results for grades in German language studies in Table 15, Appendix F).

Both regressions provide estimates of the effect of classroom gender composition that are in line with the estimates of our main specification in Table 3. For girls in STEM, math grades decrease by between 8 and 10 percent of a standard deviation for a one standard deviation increase in the proportion of female peers in the classroom (between 9 and 12 percent, depending on the measure). For girls in modern languages, math grades increase by between 6 and 8 percent of a standard deviation for a one standard deviation increase in the proportion of female peers (between 18 and 20 percent, depending on the measure). These results show the robustness of our results to different sources of variation in classroom gender composition.

6.4 Prior math grades

One potential concern is that the differential self-selection of girls into specializations based on ability might drive the effect heterogeneity that we observe across specializations. Specifically, if only the highest-ability girls in each cohort self-select into the STEM specialization, then the average ability of the girls in STEM will be higher in cohorts with fewer girls. At the same time, the average ability of girls in modern languages will be higher in cohorts with more girls. This differential self-selection of girls into specializations according

to their ability could generate a spurious correlation between the proportion of girls and math grades, a correlation that would be negative for girls in STEM and positive for girls in modern languages.

To test the robustness of our results with respect to the differential self-selection of girls into specializations, we run the same regression as in Table 3 on the subset of students that we observe prior to randomization (roughly 55 percent of the full sample), including their math grades prior to randomization as a control for ability. Table 8 reports the results of this regression. After the inclusion of math grades prior to randomization, compared to the main estimates in Table 3, the estimated coefficients on the proportion of female peers for girls in modern languages and in STEM barely change. However, controlling for prior grades does affect the coefficients for boys in modern and ancient languages.

For boys in modern languages, the difference between the coefficient in Table 8 and the main estimate in Table 3 is driven by differences in the estimation sample. For boys in ancient languages, the negative coefficient in Table 8 is driven by the presence of few mixed-specialization classrooms (ancient languages and modern languages).[7] Indeed, when we include a control variable for mixed classrooms, the estimated coefficient is no longer significant. Overall, these results indicate that the differential self-selection of girls into specializations according to their ability does not explain the heterogeneity across specializations.

7 Conclusion

In this paper, we investigate the effects of classroom gender composition on the grades of students with different preferences for the STEM fields. We use unique internal school records from one large high school in Switzerland in which students reveal their preferences for the STEM fields by selecting into a specialization (STEM vs. modern languages vs. ancient languages). By using a school policy that randomly assigns students to classrooms after they have chosen a specialization, we are able to identify the causal effect of classroom gender composition on students who self-selected into different specializations, i.e., with different preferences for STEM.

[7] Boys in ancient languages have, on average, lower math grades than girls in the same specialization, but higher math grades than girls in modern languages. Therefore, for boys in ancient languages, an increase in the proportion of female peers specializing in ancient languages implies an increase in the average math grades in the classroom, whereas an increase in the proportion of female peers specializing in modern languages implies a reduction in the average math grades in the classroom.

Although earlier research shows that a higher proportion of female peers has a positive average effect on the math grades of boys and girls, we find that this average effect masks a high degree of heterogeneity. Indeed, this effect varies systematically across students in different specializations and may even become negative for some types of students. Our most counterintuitive finding is that only girls who revealed a preference for languages (and not STEM) benefit from having more girls in the classroom. In contrast, and surprisingly, the effect of having more girls in the classroom is negative for girls who revealed a preference for STEM (but not languages).

Our analysis of additional survey data suggests that differences in the willingness to compete help explain the contrasting effects of classroom gender composition on girls in languages and STEM specializations. Future research should thus further explore the role of competitive behavior in shaping the effect of classroom gender composition.

Policy decisions regarding the gender composition of classrooms need to factor in the effect heterogeneity that we find. Decisions that consider only average effects might not yield the desired results or might even yield the opposite result. In the case of girls with a preference for STEM, average effects point in the wrong direction and might hinder a potential STEM career at a very early stage.

Table 7: Proportion of newcomer girls

	Modern languages		STEM	
	boys	girls	boys	girls
Panel A: Proportion of newcomer girls arriving in the 9th grade				
Proportion of newcomer girls	0.057	0.064*	−0.035	−0.105**
	(0.064)	(0.033)	(0.031)	(0.045)
Adj.-R2	0.084	0.053	0.016	0.061
Panel B: Proportion of newcomer girls arriving in the 9th grade and later				
Proportion of newcomer girls	0.111**	0.085***	0.011	−0.076*
	(0.053)	(0.027)	(0.029)	(0.042)
Adj.-R2	0.088	0.056	0.015	0.058
Semester FE	*yes*	*yes*	*yes*	*yes*
Teacher FE	*yes*	*yes*	*yes*	*yes*
Controls	*yes*	*yes*	*yes*	*yes*
Obs.	1163	3252	2780	1705
Ind.	233	611	555	319

Note: The dependent variable is math grade. Each coefficient represents a separate regression for the subsamples of boys and girls in different specializations. Standard errors are in parentheses. Standard errors are clustered at the classroom level. Each observation is a student in a semester. Grades are standardized by specialization and gender. The proportion of newcomer girls is standardized by specialization: the estimates report the effect (in standard deviations) of a one standard deviation increase in the proportion of newcomer girls. Controls: class size. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Prior math grades

	Modern Languages		Ancient Languages		STEM	
	boys	girls	boys	girls	boys	girls
Panel A: math						
Proportion of female peers	−0.040	0.097**	−0.167*	0.144***	0.008	−0.126***
	(0.112)	(0.038)	(0.084)	(0.017)	(0.051)	(0.044)
Prior math grade	0.702***	0.709***	0.834***	0.848***	0.871***	0.953***
	(0.127)	(0.048)	(0.077)	(0.062)	(0.101)	(0.093)
Semester FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Teacher FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Controls	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Obs.	499	1231	893	1659	1471	895
Ind.	96	221	148	268	264	151
Adj.-R2	0.312	0.336	0.379	0.459	0.262	0.382

Note: Each coefficient represents a separate regression for the subsamples of boys and girls in different specializations. Standard errors are in parentheses. Standard errors are clustered at the classroom level. Each observation is a student in a semester. Grades are standardized by specialization and gender. The proportion of female peers (leave-one-out mean) is standardized by specialization: the estimates report the effect (in standard deviations) of a one standard deviation increase in the proportion of female peers. Controls: class size, math grades prior to randomization. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

Anelli, M. and Peri, G. (2017). The effects of high school peers' gender on college major, college performance and income. The Economic Journal, 129(618):553–602.

Balestra, S., Eugster, B., and Liebert, H. (2016). Class composition, special needs students, and peers' achievement. CESifo Working Paper Series No. 6837.

Bertrand, M., Goldin, C., and Katz, L. F. (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. American Economic Journal: Applied Economics, 2:228–255.

Black, S. E., Devereaux, P. J., and Salvanes, K. G. (2013). Under pressure? the effect of peers on outcomes of young adults. Journal of Labor Economics, 31(1):119–153.

Booth, A. L., Cardona-Sosa, L., and Nolen, P. (2014). Gender differences in risk aversion: Do single-sex environments affect their development? Journal of Economic Behavior & Organization, 99(C):126–154.

Brenøe, A. A. and Zölitz, U. (forthcoming). Exposure to more female peers widens the gender gap in stem participation. Journal of Labor Economics.

Bru, E. (2006). Factors associated with disruptive behaviour in the classroom. Scandinavian Journal of Educational Research, 50(1):23–43.

Buechel, B., Mechtenberg, L., and Petersen, J. (2018). If i can do it, so can you! peer effects on perseverance. Journal of Economic Behavior & Organization, 155(C):301–314.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. The Quarterly Journal of Economics, 129(3):1409–1447.

Buser, T., Peter, N., and Wolter, S. (2017). Gender, competitiveness, and study choices in high school: Evidence from switzerland. American Economic Review (Papers & Proceedings), 107(5):125–30.

Carrell, S. E. and West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. Journal of Political Economy, 118(3):409–432.

Cools, A., Fernández, R., and Patacchini, E. (2019). Girls, boys, and high achievers. CEPR Discussion Paper 13754.

Dustmann, C., Ku, H., and Kwak, D. W. (2018). Why are single-sex schools successful? Labour Economics, 54(C):79–99.

Eisenkopf, G., Hessami, Z., Fischbacher, U., and Ursprung, H. W. (2015). Academic performance and single-sex schooling: Evidence from a natural experiment in switzerland. Journal of Economic Behavior & Organization, 115:123–143.

Fischer, S. (2017). The downside of good peers: How classroom composition differentially affects men's and women's stem persistence. Labour Economics, 46(C):211–226.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. The Quarterly Journal of Economics, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. American Economic Review (Papers & Proceedings), 94(2):377–381.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. Science, 320(5880):1164–1165.

Hill, A. J. (2017). The positive influence of female college students on their male peers. Labour Economics, 44:151–160.

Hoxby, C. (2000). Peer effects in the classroom: Learning from gender and race variation. NBER Working Papers 7867, National Bureau of Economic Research, Inc.

Jackson, C. K. (2012). Single-sex schools, student achievement, and course selection: Evidence from rule-based student assignments in trinidad and tobago. Journal of Public Economics, 96(1):173–187.

Lavy, V. and Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. American Economic Journal: Applied Economics, 3:1–33.

Lazear, E. P. (2001). Educational production. The Quarterly Journal of Economics, 116(3):777–803.

Lazear, E. P., Malmendier, U., and Weber, R. (2012). Sorting in experiments with application to social preferences. American Economic Journal: Applied Economics, 4(1):136–63.

Mouganie, P. and Wang, Y. (2019). High-performing peers and female stem choices in school. IZA Working Paper DP No. 12455.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? The Quarterly Journal of Economics, 129(3):1409–1447.

Oosterbeek, H. and Van Ewijk, R. (2014). Gender peer effects in university: Evidence from a randomized experiment. Economics of Education Review, 38(C):51–63.

Patacchini, E., Rainone, E., and Zenou, Y. (2017). Heterogeneous peer effects in education. Journal of Economic Behavior & Organization, 134(C):190–227.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Applied Statistics, 43(3):429–467.

Schneeweis, N. and Zweimüller, M. (2012). Girls, girls, girls: Gender composition and female school choice. Economics of Education Review, 31:482–500.

van Veldhuizen, R. (2017). Gender differences in tournament choices: Risk preferences, overconfidence or competitiveness? Rationality and Competition Discussion Paper Series 14, CRC TRR 190 Rationality and Competition.

Whitmore, D. (2005). Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. American Economic Review (Papers & Proceedings), 95(2):199–203.

Wolter, S. C., Cattaneo, M. A., Denzler, S., Diem, A., Grossenbacher, S., Hof, S., and Oggenfuss, C. (2014). Swiss education report 2014. Technical report, Swiss Coordination Centre for Research in Education.

Zölitz, U. and Feld, J. (2017). The effect of peer gender on major choice. Working Paper No. 270, University of Zurich, Zurich.

Appendix

A Additional tests of the validity of the identification strategy

In this section, we provide two additional tests of the validity of our identification strategy in addition to the test in section 5.1. First, we repeat the test in Table 2 for the fraction of newcomer students in 9th grade, i.e., the fraction of students who joined the high school in 9th grade for the first time. Specifically, for each student in 9th grade, we construct the binary variable "newcomer" indicating whether the student is a newcomer or was already in the school in 8th grade. Similar to grades in math and German language studies, we regress this binary variable on a battery of classroom fixed effects within specialization-by-cohort cells to test whether classrooms differ in terms of the fraction of newcomer students.

Because newcomers in 9th grade can, in principle, choose only the modern language and STEM specializations, we exclude all ten ancient language specialization cells. Moreover, we exclude the first cohort of students because we do not observe this cohort prior to randomization, and we therefore do not know whether these students were already in the school in 8th grade. In total, we run the test on 18 specialization-by-cohort cells. Table 9 reports the results of this test and shows that the number of specialization-by-cohort cells in which we detect differences across classrooms closely matches the number of expected cells under randomization.

Table 9: Test for randomness of newcomer students

	Number of specialization-by-cohort cells	
Significance level	Newcomer	Expected
1%	1	0.18
5%	2	0.9
10%	2	1.8

Note: Columns 2 and 3 report the number of specialization-by-cohort cells (out of 18) for which the test detects differences across classrooms with respect to the fraction of newcomer students in 9th grade. Column 3, "Expected", reports the number of cells for which we would expect the test to detect differences under random assignment at different significance levels.

Second, following Carrell and West (2010), we employ resampling techniques. For each 9th-grade classroom that the school created by random assignment of students, we resample 10,000 classrooms (without replacement of students) of the same size from the pool of students in the same specialization-by-cohort cell (i.e., the pool of students that was randomly assigned to classrooms in a given cohort and specialization). For each randomly sampled classroom, we compute the sum of the grades in math and German language studies prior to randomization and the average proportion of girls in the classroom. We then compute empirical p-values for each classroom as the proportion of simulated classrooms with a sum of the grades and a proportion of girls lower than the one observed in reality. If assignment is independent of prior grades and gender, the expected distribution of p-values should be uniform within specialization-by-cohort cells. We test the uniformity assumption using a one-sample Kolmogorov-Smirnov test.

For all three variables (grades in math and German language studies prior to randomization and the proportion of girls in the classroom), we reject the hypothesis of uniformity in 0 specialization-by-cohort cells. These results support our argument that conditional on the chosen specialization, the assignment of students to classrooms is independent of individual ability and gender.[8]

[8] As the Kolmogorov-Smirnov test is likely to overestimate the p-values in small samples, the results of this second test should be interpreted cautiously.

B Estimation in the full sample and test for differences in the coefficients

Table 10: Full sample

	Math				German language studies			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Proportion of female peers	0.035**	0.036*	0.030***	0.059***	−0.004	−0.006	0.011	0.003
	(0.018)	(0.018)	(0.011)	(0.019)	(0.012)	(0.012)	(0.008)	(0.014)
Ancient languages × Proportion of female peers				−0.009				0.013
				(0.022)				(0.022)
STEM × Proportion of female peers				−0.069***				0.012
				(0.023)				(0.018)
Constant	4.101***	4.257***	4.523***	4.563***	4.357***	4.187***	4.077***	4.068***
	(0.035)	(0.141)	(0.146)	(0.128)	(0.025)	(0.091)	(0.068)	(0.067)
Specialization FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Semester FE	*no*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*	*yes*
Teacher FE	*no*	*no*	*yes*	*yes*	*no*	*no*	*yes*	*yes*
Controls	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Obs.	12328	12328	12328	12328	12328	12328	12328	12328
Ind.	2279	2279	2279	2279	2279	2279	2279	2279
Adj.-R2	0.011	0.011	0.034	0.035	0.057	0.066	0.098	0.098

Standard errors are in parentheses. Standard errors are clustered at the classroom level. Each observation is a student in a semester. The coefficient on the proportion of female peers is standardized by specialization: the estimates report the effect (in standard deviations) of a one standard deviation increase in the proportion of female peers. Controls: gender, class size. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 11: Test for differences in the coefficients of Table 3

		Modern languages		Ancient languages		STEM	
		boys	girls	boys	girls	boys	girls
Modern languages	boys						
	girls	0.07					
Ancient languages	boys	0.00	0.12				
	girls	0.05	0.89	0.07			
STEM	boys	0.00	0.02	0.93	0.02		
	girls	0.00	0.00	0.10	0.00	0.01	

Note: Each cell reports the p-value associated with the null hypothesis that the difference between the two coefficients is zero.

C Exclusion of single-sex and mixed-specialization classrooms

In the test in Table 2, we excluded six classrooms that were either single-sex or mixed-specialization and that generated observable differences in the classroom gender composition within some specialization-by-cohort cells. To ensure that these classrooms do not drive our results, we run the regression in Eq. 1 in two subsamples. In the first subsample (Table 12), we exclude the six single-sex and mixed-specialization classrooms. In the second subsample (Table 13), we exclude all classrooms in specialization-by-cohort cells containing these single-sex and mixed-specialization classrooms.

Because all single-sex and mixed-specialization classrooms are either in modern languages or in ancient languages, the coefficients for the students in STEM remain unchanged. For the math grades, the coefficients in Table 12 and 13 are in line with the estimates in the full sample (Table 3). For the grades in German language studies, the estimated coefficients for the students in modern languages are consistent with the estimates in the full sample. However, both the positive coefficient for girls in ancient languages and the negative coefficient for boys in ancient languages become significant.

Table 12: Subsample without single-sex and mixed-specialization classrooms

	Modern Languages		Ancient Languages		STEM	
	boys	girls	boys	girls	boys	girls
Panel A: math						
Proportion of female peers	0.169***	0.127**	0.013	0.232***	−0.001	−0.120***
	(0.049)	(0.047)	(0.057)	(0.060)	(0.027)	(0.035)
Adj.-R2	0.090	0.043	0.018	0.062	0.015	0.064
Panel B: German language studies						
Proportion of female peers	−0.018	0.087	−0.146*	0.117*	0.059	−0.110**
	(0.067)	(0.063)	(0.084)	(0.065)	(0.037)	(0.051)
Adj.-R2	0.097	0.045	0.095	0.040	0.044	0.096
Semester FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Teacher FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Controls	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Obs.	1071	2839	1240	1736	2780	1705
Students	218	540	218	305	555	319

Note: Single-sex and mixed-specialization (modern and ancient languages) classrooms are excluded. Each coefficient represents a separate regression for the subsamples of boys and girls in different specializations. Standard errors are in parentheses. Standard errors are clustered at the classroom level. Each observation is a student in a semester. Grades are standardized by specialization and gender. The proportion of female peers (leave-one-out mean) is standardized by specialization: the estimates report the effect (in standard deviations) of a one standard deviation increase in the proportion of female peers. Controls: class size. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 13: Subsample without specialization-by-cohort cells containing single-sex and mixed-specialization classrooms

	Modern Languages		Ancient Languages		STEM	
	boys	girls	boys	girls	boys	girls
Panel A: math						
Proportion of female peers	0.165***	0.181***	0.011	0.219***	−0.001	−0.120***
	(0.058)	(0.051)	(0.069)	(0.068)	(0.027)	(0.035)
Adj.-R2	0.107	0.046	0.019	0.049	0.015	0.064
Panel B: German language studies						
Proportion of female peers	0.038	0.094	−0.212**	0.151**	0.059	−0.110**
	(0.073)	(0.069)	(0.097)	(0.067)	(0.037)	(0.051)
Adj.-R2	0.091	0.053	0.097	0.039	0.044	0.096
Semester FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Teacher FE	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Controls	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Obs.	864	2320	1087	1580	2780	1705
Ind.	189	460	191	267	555	319

Note: Specialization-by-cohort cells containing single-sex and mixed-specialization (modern and ancient languages) classrooms are excluded. Each coefficient represents a separate regression for the subsamples of boys and girls in different specializations. Standard errors are in parentheses. Standard errors are clustered at the classroom level. Each observation is a student in a semester. Grades are standardized by specialization and gender. The proportion of female peers (leave-one-out mean) is standardized by specialization: the estimates report the effect (in standard deviations) of a one standard deviation increase in the proportion of female peers. Controls: class size. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D Nonlinear effects

The existence of nonlinearities potentially offers scope for efficiency gains by identifying which classrooms would benefit most from an increase in the proportion of girls.

To analyze these nonlinear effects, we run a fractional polynomial regression in the spirit of Royston and Altman (1994). This approach offers a greater degree of flexibility to model nonlinear relationships compared to standard polynomials.[9] Based on the fitted fractional polynomial regression, we derive the estimated marginal effects of the proportion of female peers on math grades for classrooms with different gender compositions. Fig. 5 shows these marginal effects.

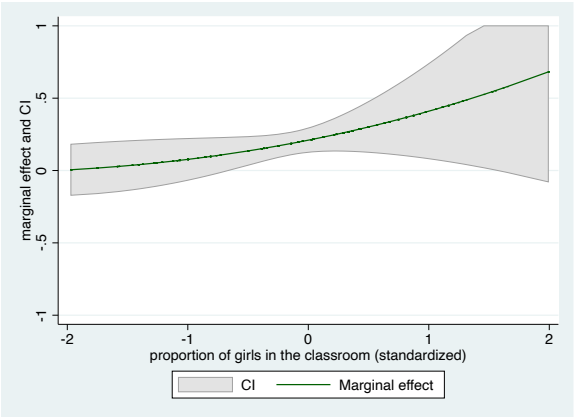
The effects appear to be nonlinear for girls in modern languages and for girls in ancient languages (Fig. 5b and 5d), and the patterns are very similar in both specializations: an increase in the proportion of female peers produces larger positive effects when there are few girls in the classroom. However, the positive effect of adding more girls to the classroom decreases with the number of girls who are already in the classroom. These findings indicate that for girls in the language specializations, the benefit of increasing the proportion of female peers is particularly large in classrooms with few girls. In contrast, the effect for girls in STEM (figure 5f) is negative and relatively flat at different proportions of girls in the classroom.

For boys in ancient languages and STEM (Fig. 5c, and 5e), the plots provide little evidence of nonlinear effects. Moreover, the confidence intervals include the null value over the full range of the proportion of female peers. For boys in modern languages (Fig. 5a), the effect appears to increase with the proportion of female peers in the classrooms. However, the estimated coefficient is precisely estimated only when the standardized proportion of female peers is roughly between -0.5 and +0.5 standard deviations.

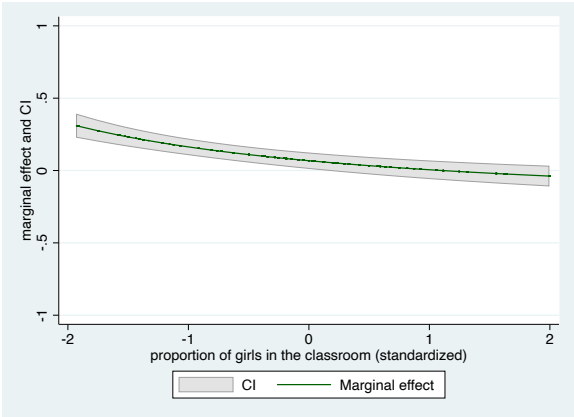
Taken together, the findings indicate that the optimal gender composition of classrooms with respect to math grades is different for different specializations. In the ancient language specialization, classrooms with a relatively high proportion of girls are efficient, as this gender composition benefits girls and leaves boys unaffected. However, the same does not apply to the STEM specialization. Indeed, STEM classrooms with a high proportion of girls are likely to be inefficient, as girls who select into STEM benefit from being in classrooms

[9] Specifically, we run regression 1 in each subsample of girls and boys in different specializations substituting $\gamma ratio_{cs}$ with a fractional polynomial of the generic form $\gamma_1 ratio_{cs}^{\rho} + \gamma_2 ratio_{cs}^{\sigma}$, where $\rho, \sigma \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. To do so, we use the Stata command *fp*.

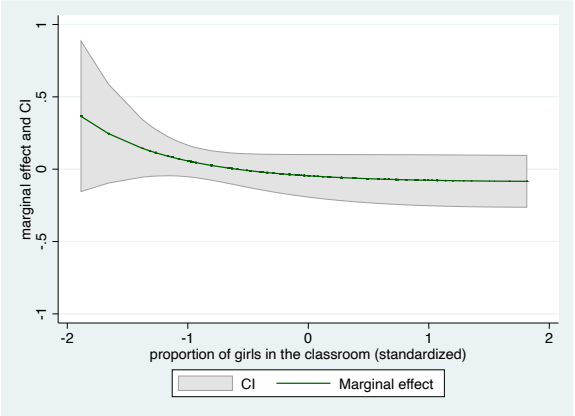
Figure 5: Nonlinear effects



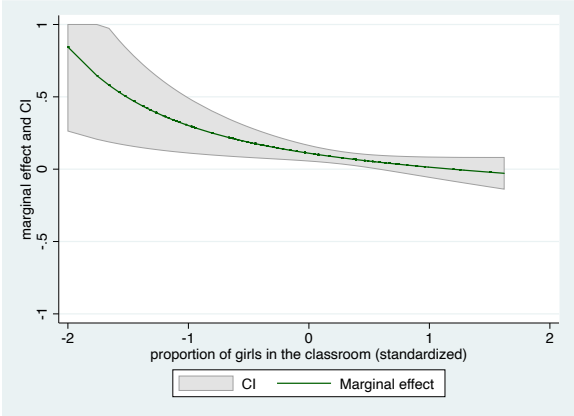
(a) Boys in modern languages



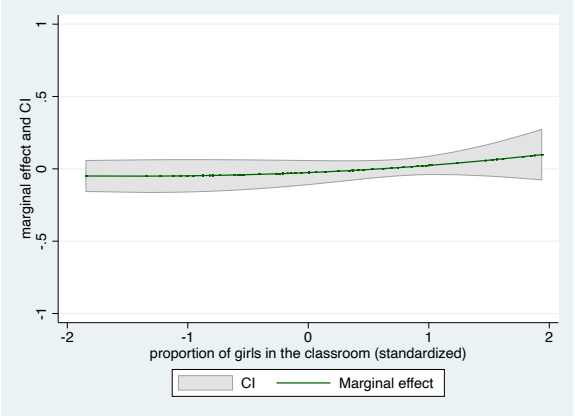
(b) Girls in modern languages



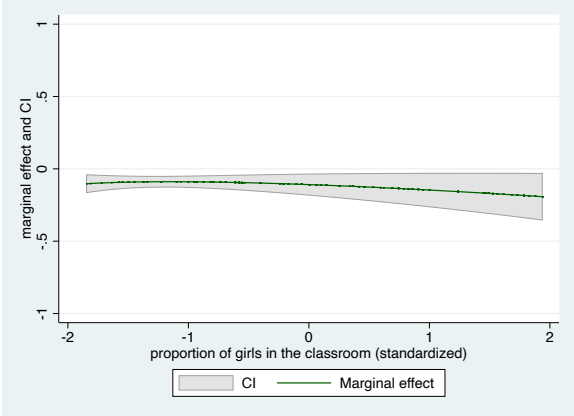
(c) Boys in ancient languages



(d) Girls in ancient languages



(e) Boys in STEM



(f) Girls in STEM

with fewer girls. In modern languages, girls and boys both benefit from a higher proportion of girls, so the total effect of moving girls across classrooms is a priori unclear. However, for a given distribution of girls across classrooms, assigning a new girl to a classroom with few girls generates larger positive effects than assigning her to a classroom with many girls.

E Self-reported student data

To complement the school's internal records, we surveyed one cohort of 12th graders (graduation in 2014) in all three specializations. The survey is divided into three sections.

The first section consists of questions about the students' interest, motivations, and attitudes towards math and German language studies. The questions are derived from the questionnaire used by Eisenkopf et al. (2015). The second section asks the students to evaluate their classroom environment, with the questions taken from the "Linzer Fragebogen zum Schul- und Klassenklima", a well-established German questionnaire for evaluating both classroom and school environments. The third section asks the students to provide personal and family socioeconomic background information. In this last section we follow Eisenkopf et al. (2015) and construct the items along the lines of the PISA questionnaire items. As in Eisenkopf et al. (2015), we further amend this part of the questionnaire by including a question on the students' self-described inclination towards competitive behavior.

A total of 177 students in all three specializations completed the questionnaire in December 2014: 80 students in modern languages, 23 in ancient languages, and 74 in STEM. Despite the low number of observations, we can exploit the self-reported student data to gain insights into differences in socioeconomic characteristics between students in different specializations. Table 14 reports the means of the socioeconomic variables from the third section of the survey and their differences in means across specializations.

Table 14: Socioeconomic variables

Item	Modern languages (N=80)	Ancient languages (N=23)	Δ Ancient languages	STEM (N=74)	Δ STEM (N=74)
1 Hours spent reading per day	2.175	2.522	0.347	2.315	0.14
			(0.243)		(0.166)
2 Number of books	4.175	4.565	0.39	4.329	0.154
			(0.319)		(0.219)
3 Number of cellphones	4.288	4.043	-0.244	4.945	0.658
			(0.385)		(0.264)
4 Number of PCs	3.6	3.391	-0.209	4.877	1.277***

			(0.481)		(0.329)
5 Number of bathrooms	2.025	1.913	-0.112	2	-0.025
			(0.167)		(0.114)
6 Number of rooms	4.861	5	0.139	5.528	0.667***
(excluding kitchen and bathrooms)			(0.33)		(0.228)
7 Number of televisions	1.775	1.739	-0.036	1.822	0.047
			(0.227)		(0.155)
8 Number of cars	1.55	1.261	-0.289	1.288	-0.262*
			(0.23)		(0.157)
9 Number of music instruments	3.09	3.957	0.867	2.589	-0.501
			(0.701)		(0.481)
10 Competitiveness	4.519	4.304	-0.215	5.801	1.282***
			(0.661)		(0.454)
11 Fraction of German native speakers	0.663	0.739	0.077	0.662	-0.001
			(0.112)		(0.076)
12 Father's education					
High-school dropout	0.057	0.091	0.034	0.028	-0.029
(less than 12 years of schooling)			(0.053)		(0.037)
High-school graduate	0.371	0.318	-0.053	0.296	-0.076
			(0.116)		(0.08)
Vocational Education and Training	0.143	0.045	-0.097	0.113	-0.03
			(0.079)		(0.054)
College	0.429	0.545	0.117	0.563	0.135
			(0.122)		(0.084)
13 Mother's education					
High-school dropout	0.111	0.136	0.025	0.042	-0.069
(less than 12 years of schooling)			(0.068)		(0.047)
High-school graduate	0.389	0.364	-0.025	0.366	-0.023
			(0.119)		(0.082)
Vocational Education and Training	0.139	0.045	-0.093	0.169	0.03
			(0.085)		(0.058)
College	0.361	0.455	0.093	0.423	0.061

			(0.12)		(0.082)
14 Fraction of students living	0.868	0.826	-0.042	0.925	0.057
in the city of Zurich (vs. suburbs)			(0.076)		(0.054)

Note: Difference in means of students in ancient languages and STEM relative to students in modern languages. Standard errors are in parentheses. Question 1 asks the students to indicate the number of hours per day they spend reading. Question 2 asks them to choose one of six options indicating the number of books that they have at home (including the books owned by other family members): 1) 0-10; 2) 11-25; 3) 26-100; 4) 101-200; 5) 201-500; 6) more than 500. Questions 3 to 9 asks the students to indicate the number of cellphones, PCs, rooms, bathrooms, televisions, cars, and music instruments they have at home (including those owned by other family members). Question 10 asks the students about their willingness to compete on a scale from 1 to 10: 1="I do not like to engage in competition at all" to 10="I really like engaging in competition". Question 11 asks students if they are German native speakers (0=no; 1=yes). Questions 12 and 13 ask about parental education. Question 14 asks about the type of municipality of residence (0=suburbs; 1=city). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

F Proportion of newcomer girls: German language studies

Table 15: Proportion of newcomer girls

	Modern languages		STEM	
	boys	girls	boys	girls
Panel A: Proportion of newcomer girls arriving in the 9th grade				
Proportion of newcomer girls	-0.158^{***}	0.025	-0.022	-0.083
	(0.044)	(0.024)	(0.048)	(0.065)
Adj.-R2	0.132	0.045	0.042	0.093
Panel B: Proportion of newcomer girls arriving in the 9th grade and later				
Proportion of newcomer girls	0.098^{**}	0.038	0.063^{*}	-0.072
	(0.042)	(0.040)	(0.037)	(0.081)
Adj.-R2	0.124	0.046	0.044	0.093
Semester FE	*yes*	*yes*	*yes*	*yes*
Teacher FE	*yes*	*yes*	*yes*	*yes*
Controls	*yes*	*yes*	*yes*	*yes*
Obs.	1163	3252	2780	1705
Ind.	233	611	555	319

Note: The dependent variable is grade in German language studies. Each coefficient represents a separate regression for the subsamples of boys and girls in different specializations. Standard errors are in parentheses. Standard errors are clustered at the classroom level. Each observation is a student in a semester. Grades are standardized by specialization and gender. The proportion of newcomer girls is standardized by specialization: the estimates report the effect (in standard deviations) of a one standard deviation increase in the proportion of newcomer girls. Controls: class size. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.