Peer Gender and Schooling

Borbely, Daniel; Norris, Jonathan; Romiti, Agnese

Published in:
Journal of Human Capital

DOI:
10.1086/723111

Publication date:
2023

Peer Gender and Schooling: Evidence from Ethiopia*

Daniel Borbely†

Jonathan Norris‡

Agnese Romiti§

May 18, 2023

Abstract

This paper studies how classmate gender composition matters for school absences and test scores in a context characterized by strong social norms and scarce school resources. We base our results on a unique survey of students across classrooms and schools in Ethiopia, exploiting random assignment of students to classrooms. We find a strong asymmetry: while females benefit from exposure to more female classmates with reduced absenteeism and improvement on math test scores, males are unaffected. We further find that exposure to more female classmates improves motivation and participation in class, and in general, that the effects of classmate gender composition are consistent with social interaction effects.

JEL classification numbers: I21, I29, J16, J24
Keywords: Peer Effects, Gender, School Performance, Ethiopia

*We thank Lukas Kiessling, Dora Gicheva, seminar participants at the University of Strathclyde and the University of North Carolina at Greensboro, and conference participants at the 34th Annual Conference of the European Society for Population Economics for many helpful comments. Declarations of interest: none.

†Corresponding author. Email: dborbely001@dundee.ac.uk. University of Dundee

‡Email: jonathan.norris@strath.ac.uk. University of Strathclyde

§Email: agnese.romiti@strath.ac.uk. University of Strathclyde

1 Introduction

In this paper, we study the effect of classroom gender composition on absence from school and test scores, using random assignment of students to classrooms in Ethiopia.[1] While there is a large literature studying the effects of peer gender composition on educational outcomes, these studies predominately use data from more developed and Western countries.[2] Peer effects in a developing context – where systems, incentives, peer groups, and norms may differ – have received less attention.[3] Furthermore, even within the current literature it is not clear whether we should expect symmetric effects across gender – for instance, females and males experiencing similar responses and mechanisms to classmate gender composition – or asymmetric effects. Asymmetry may be particularly likely where gender stereotyping is strong and females benefit especially from exposure to more females.

In the past, school enrollment in Ethiopia was a considerable problem; however, since policy reforms in the early 2000s, the country has experienced growth in enrollment at all levels of schooling and the gender gap in enrollment has narrowed (UNICEF Ethiopia, 2019). Nevertheless, there remain salient issues for children's educational progress. Constraints to education arise from late entry to school and early departure, and importantly, irregular attendance to school (Boyden et al., 2020, Favara, 2017, Tafere and Pankhurst, 2015). Social norms can be strong, often related to traditional gender roles, and have a strong influence on children's time use (Favara, 2017).[4] In addition, class size tends to be very large, establishing an environment where teachers are spread

[1] Ethiopia is a fast-growing developing country, which is ranked 12^{th} worldwide in terms of population size (World Development Indicators, 2019).

[2] See for example Hoxby (2000), Lavy and Schlosser (2011), Black et al. (2013), Cools et al. (2019).

[3] To our knowledge, Duflo et al. (2011), and Hahn et al. (2019) represent the first studies. Duflo et al. (2011) evaluate the effects of tracking by student initial achievement using experimental data from Kenya but focuses on quality of peers in terms of test score rather than gendered peers. Hahn et al. (2019), exploiting an experimental setting in Bangladesh, compares the effect of studying in groups with friends vs groups with peers on academic outcomes of school children.

[4] For instance, girls tend to engage primarily in domestic chores within the household, while boys tend to contribute to activities outside the household such as herding, farming, or paid work (Favara, 2017, Tafere and Pankhurst, 2015). Also, girls in particular have an expectation to maintain a good social reputation so they can secure a marriage once of age (Coles et al., 2015, Tafere and Chuta, 2016).

thin and peers may form an important source of influence not only on performance but also on incentives to go to school. Thus, students face significant differences in norms across gender, competition for their time, and features of the school environment that suggest peers may play a significant role.

In general, we can define potential mechanisms for peer gender effects under direct effects (social interactions) or indirect effects (e.g., shifts in teacher behavior). Within these, there exist a number of feasible mechanisms whereby classmate gender composition may affect student outcomes and either lead effects to be similar or to differ across genders. For instance, if females tend to exhibit fewer externalizing behavioral problems – such has been documented in the early development of children in the US (Bertrand and Pan, 2013) – then having more female classmates may benefit both females and males directly or indirectly through classroom mechanisms such as teachers.[5]

Alternatively, students may have beliefs about their capabilities that make them more or less confident to engage in their studies or in the classroom. For example, where gender stereotypes are negative toward girls' academic effort, then as the share of females in the class increases, social interactions may reduce the saliency of gender norms and build confidence among females. This interpretation would be consistent with an adaption of the identity model in Akerlof and Kranton (2000). Females are prescribed stereotypical gendered behavior and the cost to identity utility when deviating from gender norms increases for females in the presence of males. Recent evidence confirms that beliefs and gender stereotypes can operate as a significant mechanism by lowering females' beliefs about their ability (Bordalo et al., 2019), and females' likelihood to engage in competition (Niederle et al., 2013, Booth and Nolen, 2012). Where beliefs and gender stereotyping are the dominant mechanism, then we would expect class gender composition effects to be focused among girls.

Peer gender effects may also arise from other sources. First, increases in the share of female peers could act to protect girls from bullying, or hostile environments, in the classroom. In this

[5] Lavy and Schlosser (2011) also find a reduction in classroom disruption or violence and improved teacher-student relations consistent with this potential mechanism.

case, effects may again be focused on girls. This phenomenon may be particularly relevant for the Ethiopian context, where there is no legislation or policies targeted to tackle bullying (Pells et al., 2016), even though evidence from developed countries shows bullying has detrimental effects on educational attainment and future earnings (Brown and Taylor, 2008, Eriksen et al., 2014). Second, whether indirect effects via teachers would generate symmetric or asymmetric effects depends on the context. For example, if more girls in the class improves teacher bias, or stereotypes, toward girls, then girls may benefit from better attention from teachers. This indirect channel could lead to asymmetry. Thus, a number of mechanisms for peer gender effects suggest it is important to assess the role of heterogeneity across gender.

The main contribution of our study is twofold. First, we provide evidence on peer effects in an environment very different from the existing evidence base. This is in terms of the potential role for female peers due to fewer school resources, as well as the strong gender norms that exist in Ethiopia (Favara, 2017). Compared to, for example, Gong et al. (2021), who study peer gender effects in China, the gender gap in school enrollment as well as the average class size are considerably lower in China than in Ethiopia.[6] The Ethiopian context represents a considerably different context where gender peer effects may operate. Second, we provide new evidence on the role of classmate gender composition and mechanisms that can drive asymmetric effects across gender.

Our paper also relates more broadly to a literature evaluating how features of school environments affect student outcomes. These features include the consequences of class size (Angrist and Lavy, 1999, Krueger and Whitmore, 2001, Chetty et al., 2011, Angrist et al., 2019), teacher quality (Chetty et al., 2014, Rothstein, 2017), effects of tracking by initial achievement conditional on teacher incentives (Duflo et al., 2011), and peer effects over a range of dimensions. These peer dimensions include long-term effects of exposure to disruptive peers on earnings (Carrell et al., 2018), the link between exposure to low-achieving Kindergarten peers and non-cognitive skills

[6]In China, there is no difference in gross enrolment ratio (96%) across gender (World Bank, 2015), whereas in Ethiopia gross enrollment ratio is much higher for boys (106% boys vs 96% for girls). In addition, average class size in primary school is 37 in China (OECD, 2014) compared to 60 in Ethiopia (see Table 1).

(Bietenbeck, 2020) and between academic achievement and peers' persistence (Golsteyn et al., 2020), spillovers from friends' educational aspirations (Norris, 2020, Gagete-Miranda, 2020), and the effects of a variety of peer compositions on educational attainment.[7] We add to the literature on peer effects by assessing exposure to the share of female classmates within a new context where classes are large, teachers likely spread thin, and gender norms are strong.

We empirically test the effects of classroom gender composition on educational outcomes across genders using data from the Young Lives Ethiopia school survey. This data was collected from in-school surveys and administered tests at the beginning and end of the 2012-13 school year for students in selected 4th and 5th grade classrooms. We then leverage information for each classroom from teachers on whether students were randomly assigned and assess causal effects from the share of female classmates.[8]

By exploiting random assignment of students to classrooms, we find that an increase in the share of female classmates decreases missed school and raises math test scores for females, while having no effect on males. Among females, we find that classmate gender composition is an important feature of the school environment in Ethiopia: a standard deviation (9 percentage point) shift in the share of female peers translates into approximately one day less of missed school and 7% of a standard deviation increase in math test scores. Moreover, our results remain stable through a range of robustness checks. We also assess additional heterogeneities around factors that may capture individual disadvantage, moderating effects from school or class characteristics, and nonlinearities. In general, we find very little evidence of substantial heterogeneities along these dimensions, though we find suggestive evidence that the benefits from female peers are stronger as their share increases.

We then assess evidence around mechanisms. We focus on evidence supportive of either social interactions or shifts in teacher behaviour. First, we find that females and males experience improvements in motivation and class participation as a result of sharing the classroom with more

[7] For instance, these include the effect of immigrant school-grade composition (Gould et al., 2009) and peers' parents education (Bifulco et al., 2011, 2014, Fruehwirth and Gagete-Miranda, 2019).

[8] In Section 4.2, we show that a broad range of balance tests are highly consistent with expectations given random assignment. Furthermore, our robustness checks in Section 5.2 indicate very little sensitivity in our results.

female peers. While these effects are symmetric, they suggest that girls indeed become more motivated and participate more in the presence of more girls, which can then serve as a channel to boost their attendance and performance.

Second, our analysis shows that teacher behaviors and attitudes do not change in response to the share of females in the class. These teacher outcomes cannot cover all possible types of teacher responses to shifts in the gender composition, also some of these measures are taken a short time after the treatment. Thus, while our evidence points away from shifts in teacher behaviors as the mechanism, we temper this conclusion and present it as suggestive evidence in-line with social interactions as the driving mechanism.

Third, we show that for school absences the positive effect among girls tends to be smaller when boys in the class are older, regardless of the average age of girls. Conversely, for math scores, boys' age does not matter for the positive effect girls experience, while the age of girls in the class does.[9] In Section 5.4.3, we discuss these results and argue that the patterns on both outcomes along classmate age are consistent with social interaction effects. While we cannot directly test for gender stereotyping and beliefs, our results are at least highly consistent with mechanisms that boost motivation and allow girls to perform up to their ability.

Finally, we investigate the moderating influence of child work on the peer effect from having a higher share of girls in the classroom. In the developing context, the presence of child work can offset positive early educational influences (Bau et al., 2020). This issue is particularly pertinent in Ethiopia, where child work is widespread, and schoolchildren often have to balance schooling with work commitments and domestic chores. We find that the presence of child work does moderate the effect of peer gender on school absences and math scores. Importantly, however, girls engaged in a high amount of child work still benefit from an increased share of female classmates, suggesting that peers can form a strong source of influence even in the presence of detrimental educational environments.

[9]We find stronger, positive effects for females on math scores where female peers are not aged in top tertile of the female peer age distribution.

We find persistent evidence that classmate gender composition impacts important educational outcomes. Females drive the effect and experience strong, positive effects from exposure to more females in the classroom. Our results are consistent with mechanisms driven by social interactions and, while not proving, add support for models to incorporate beliefs and gendered norms in the production of skills. Moreover, our results further show peers can be an important source of influence within a developing context, where effects are likely shaped by the class environment.

2 Ethiopia: Education and Institutional Background

The Ethiopian context is very different from the US and European settings where much of the analysis on peer gender composition has taken place. The majority of the country's population resides in rural areas, where the provision of primary education is made more difficult by a dispersed population, poor infrastructure, and political instability. Primary school enrolment rates have increased from 20% in 1991 to 85% in 2011 due to large-scale educational expansion and school-building programs implemented by the Ethiopian Government (Orkin, 2013). The rapid expansion of the primary education system nonetheless came at the expense of school quality, which remained low in many areas due to teacher shortages, high pupil-teacher ratios, and poorly built schools.

The current education system in Ethiopia was established through the 1994 Education and Training Policy. Formal education begins at age 7 with primary school. This lasts from Grade 1 to Grade 8 (the first cycle is between grades 1–4, the second cycle is grades 5–8) followed by secondary education through Grades 9–12 (where the last two grades are for university preparation). Exams are taken at Grades 8, 10 and 12. The regional exams taken at Grade 8 certify the completion of primary school education (Tafere and Tiumelissan, 2020).

Students typically attend school five days a week for 39 weeks per year. Each school day is four hours divided into six periods of 40 minutes (Ministry of Education, 2009a). Out-of-school children account for 14% of all primary school aged children in the country, but this average figure

masks large regional disparities — the share of primary school aged children out of school is 1.1% in Addis Ababa but 59.6% in Afar (UNICEF Ethiopia, 2019).

Students tend to progress through school relatively slowly in Ethiopia, as repeating grades and dropping out of school are common even during primary school. In 2016/17, the primary school completion rate (finishing Grade 8) was only 54.1%, 56% for boys and 52.2% for girls (Tafere and Tiumelissan, 2020). As a result of students often repeating grades, a high proportion of children in primary schools are over-age. The main causes of school interruptions tend to be child work, poverty, illness, or lack of interest in school due to poor teaching quality (Tafere and Pankhurst, 2015, Tafere and Tiumelissan, 2020, UNICEF Ethiopia, 2019). Strong gender norms also play a role in school interruptions, as boys are likely to miss school due to being engaged in activities such as herding, farming, or paid work, while girls are likely to be absent due to domestic chores or family commitments (Tafere and Pankhurst, 2015, Favara, 2017).

Poor teacher incentives, absenteeism and low teaching quality are common impediments to effective schooling in developing countries (Kremer and Holla, 2009). Qualitative evidence indicates that these issues are also pertinent in the Ethiopian school system, and particularly in rural schools (Abebe and Woldehanna, 2013, Tafere and Pankhurst, 2015, Tafere and Tiumelissan, 2020). Teacher absenteeism in Ethiopian schools is mostly driven by factors such as teacher shortages, poor teacher incentives and compensation, inadequate management of teachers and schools by headteachers, and the lack of appropriate teaching facilities and infrastructure (Yadete, 2012, Abebe and Woldehanna, 2013).

Overall, issues surrounding school and teacher quality, along with the presence of gender norms and markedly different educational and life trajectories for boys and girls, could make peer gender effects a particularly salient channel for educational improvements in the Ethopian context.

3 Data

3.1 Young Lives Ethiopia School Survey

We use data from the Young Lives Ethiopia school survey covering the 2012-2013 school year. School sites were selected across 30 locations within Ethiopia with all schools within the location included. In the full sample, there are 92 schools and 280 classrooms. The survey is not representative of the population but it was designed to capture a wide range of environments within the country (Aurino et al., 2014). The survey covers five out of the nine Ethiopian regions, where more than 96 percent of the population lives: Addis Ababa, Amhara, Oromia, SNNP, and Tigray. In each region, between three and five *woredas* (districts) were selected (20 in total), with a balanced sample of households from different parts of the income distribution, from different types of urban and rural areas alike. Within each *woreda* at least one *kebele* (local administrative area) was chosen for the sample. In general, the YL data oversamples poor households, while the school survey we rely on has a larger than national share of students in schools in urban areas (81% in our final sample), partly due to urban schools simply having more students.

Two waves of survey collection occurred, including all grade 4 and 5 classes within a school and all students enrolled in one of these classes who were present on the day of the survey. The first survey was conducted near the beginning of the school year with nearly 12000 students. This includes a grade appropriate math test, a literacy test, and questionnaires from students, teachers, and school principals.[10] The second survey was conducted near the end of the school year.[11]

The math and literacy tests were re-administered, and this survey includes updated information on the pupil, class, and teacher rosters, including information for each student on days of missed

[10]The math and literacy tests were given in the language of instruction used in the class and supervised by the Young Lives fieldworkers.

[11]Students who left the school are not followed. Only students included at the start-year survey and who are present at the end-year survey collection are included (Aurino et al., 2014). Of the 11591 students with valid math and literacy start-year test scores, 9777 (or 84.4%) complete both math and literacy end-year tests. In Section 5.2, we examine whether attrition is related to gender class composition and show that our baseline results are robust to correcting for attrition.

school. The numbers of days absent for each student is from the class roster, which is filled in by teachers at the end of the school year. Also, included for each student are motivation and class participation scales reported by the teachers that we use in our mechanisms section.

3.2 Sample Selection

We focus on end-year outcomes that are related to the production of skills: absences and test scores. Absences are reported for each student in the class roster as the number of days absent since the start-year survey. Math and language test scores, at both the start-year and end-year, each consist of 25 items.[12] For each item, we observe whether the student gave the correct answer and from these construct item response theory (IRT) scores. IRT scores provide consistent measures of latent math and language ability that we can compare across age groups (see Van Der Linden and Hambleton, 1997). The IRT model assumes that each multiple choice item on a test is characterised by an Item Characteristic Curve (ICC). The ICC then maps each student's latent ability into the probability that they answer a particular question correctly.[13]

Peer variables are constructed from the start-year survey at the class level as leave-one-out means. Our focus, or peer treatment, is the leave-one-out mean share of female classmates. We also construct peer means for start-year test scores and for each of our student characteristic controls.

Empirically, we aim to analyze the causal effect of classmate gender composition. Thus, we leverage information from the class level portion of the survey on the method of student assignment to the classroom. At the start-year sample, we restrict the data to classrooms reporting random assignment (8234 observations). The survey indicates whether students were assigned to a class

[12] We refer to the literacy test as the language test score throughout the paper.

[13] In this study, we primarily use the standard two-parameter IRT model, which does not account for the correct guessing of answers. Our results do not change when we calculate math IRT scores using the three-parameter model, which factors in the probability that a student correctly guesses an answer. Due to lack of convergence, unfortunately, we are unable to calculate three-parameter IRT scores for our language test score variable.

"randomly/alphabetically".[14] Where assignment is alphabetical, then a concern is whether there could be clustering of students with similar last names based on ethno-linguistic or ethno-religious characteristics. In Section 4.2, we provide evidence that classroom assignment is random in our sample through a series of balance checks. Additionally, language in Ethiopia likely captures ethno-religious differences (Ado et al., 2021). In later robustness checks, we include home-language fixed effects and find no sensitivity in our results (see Section 5.2). Moreover, Figure A.1 in the Appendix shows that home language does not predict classroom gender composition in our sample.

As we always include school fixed effects, we drop observations in schools with less than 2 classrooms.[15] Further, we drop those missing on key start-year variables, i.e. gender, the share of female classmates, and class size.[16] We then drop those missing end-year days absent and test score outcomes. Of those present in our start-year selected sample, 16.7% (1117 observations) do not record end-year math and language tests.[17]

Next, because the share of female classmates is the focus of our analysis, we use the Fisher's exact test to evaluate whether classroom gender ratios are consistent with random assignment in our sample. If our sample equates to a randomly assigned sample, then gender distribution should be roughly similar across classrooms within a school to the overall gender ratio at the school-level. We keep observations in schools that fail to reject the null that gender is randomly assigned across classrooms, with a *p-value* larger than 0.10. In the sample described above, 92% of the data pass this test. Our final selected sample size contains 5077 observations across 41 schools and 132 classrooms.

[14]The full list of categories for student assignment into classrooms indicated in the YL survey are: randomly/alphabetically, by ability, by language, by gender, by age, or other method.

[15]Within the sample reporting random allocation to classrooms, this amounts to only 166 (2% of random allocation sample) observations. Moreover, this also leaves no classroom reporting fewer than 22 observed students. We also drop a small number of observations for whom the reported class size is less than the calculated observed class size. These amount to 194 observations or 2.4% of the sample reporting random allocation to classrooms.

[16]There are 6676 observations after these steps in our base start-year sample.

[17]We only lose 14 more observations who are in our base start-year sample and have valid end-year test scores but are missing information for end-year days absent. After these steps, we are left with 5545 observations.

3.3 Summary Statistics

In Table 1, we report summary statistics broken down by gender for our baseline set of outcomes, key variables, and controls in the selected sample. The last column reports for each variable the difference between boys and girls with the respective significance level. There are some significant differences between boys and girls, however most of them are small, and there is no systematic difference pointing to a specific direction. For example, boys have slightly higher absenteeism, score worse in terms of language both at the beginning and end of the year, but slightly better in start-year math test scores. On the other hand, it is more likely for boys to belong to a family where a minority language is spoken at home, or to have an illiterate mother, whereas boys seem to be more likely to live with the fathers and their biological mother. In addition, we believe that it is unlikely that differences in start-year test scores are driving the results as they mostly point to differences in language that is not affected by our treatment. Interestingly, the quality of start-year peers in terms of both language and math is worse for boys. However, these descriptives are unconditional on school fixed effects, which is the level at which classmate peers can be drawn, and our identifying assumption relies upon. Comparing selected and non-selected sample in the Appendix Table A.1, on average, students have missed nearly 6 days of school by the end-year survey – the mean masks significant variation with a standard deviation of about 7 days.[18] Average test scores in the selected sample are higher than the mean in the full sample – at both the start and end-year surveys. In Table A.1, we also show that means for our outcomes are statistically different between the selected and non-selected sample with days absent smaller (0.87 fewer days) and test scores larger in the selected sample, suggesting some degree of positive sample selection. Mean gender is statistically the same across the selected and non-selected samples, while the remaining characteristics are statistically different. However, in all cases, these differences are small and do not represent a clear pattern of

[18] Information on absences is reported in the class roster, which is completed by a fieldworker through the class teacher and school principal. Though this measure is not self-reported, our results show no correlation between reporting zero absences at the end of the year and our peer gender variable (results available upon request).

advantage or disadvantage.[19]

In Appendix Figure A.2, we report histograms for the share of female classmates within the selected sample. We show both the raw variation and variation post-removal of school fixed effects. There is considerable support with nearly continuous variation that is approximately normally distributed and ranges from 35% to 65% of the percent of female peers, suggesting sufficient variation to identify our effects of interest. We also report this variation split across quartiles of class size (see Appendix Figure A.3) out of concern that our variation is driven only by smaller classes. This however, is not the case, with variation in the share of female classmates rather similar across the distribution of class size. Gong et al. (2021) is another study exploiting random assignment of students to classrooms in a non-western context (China), which reports the same raw variation as our study in the share of female classmates.[20]

The remaining variables represent the controls that we include for student characteristics and class-level characteristics. The sample is evenly split by gender. Average age is approximately 11.5 years and average age at school start near 6.7 years. Aurino et al. (2014) note students in the full sample are on average in the appropriate age range for the surveyed grades but that there is heterogeneity in this, stemming from late school starters. Therefore, in all specifications, we flexibly control for both age and age at school start with quadratics.

In Ethiopia, there are a large number of languages that students respond with as their language spoken at home, with the majority speaking Amharic. We include a simple indicator for speaking a minority language at home in all specifications.[21] The remaining controls capture characteristics about the household – number of older and younger siblings – and about parents – having both

[19]For instance, mean age is 11.55 in the selected sample and 11.45 in the non-selected sample, suggesting the selected sample is slightly older, but mean paternal literacy is 50% for mothers and 57% for fathers in the selected sample compared to 46% and 60% in the non-selected sample. Our selected and non-selected samples have a similar proportion of private school students as well, alleviating potential concerns over the classroom assignment mechanism being different for public and private schools. In addition, the distribution of pupils across private and public schools in our sample matches figures from official statistics (Ministry of Education, 2009b).

[20]Table 1 in their paper reports a raw standard deviation of 0.08, virtually identical to what we find. Despite the Chinese context may well be very different from the context we are examining, not least because classes are much smaller, we are not aware of any other study we could use as a way of comparison.

[21]We do relax this in later robustness checks by including home language fixed effects.

Table 1. Summary Statistics

	Female		Male		Difference
	Mean	SD	Mean	SD	Diff
Outcomes					
End-Year Days Absent	5.25	6.71	5.79	7.89	0.54**
End-Year Math Test Score	-0.02	0.98	0.02	1.02	0.03
End-Year Language Test Score	0.07	0.99	-0.08	1.00	-0.15***
Peer Variables					
Share Female Peers	0.51	0.08	0.49	0.09	-0.02***
Peer Start-Year Math Scores	0.10	0.47	0.07	0.48	-0.02
Peer Start-Year Language Scores	0.08	0.61	0.02	0.62	-0.05**
Start-Year Test Scores					
Own Start-Year Math Scores	0.08	0.87	0.13	0.87	0.05*
Own Start-Year Language Scores	0.12	0.92	0.02	0.90	-0.10***
Student Characteristics					
Age (years)	11.48	1.60	11.63	1.60	0.15***
Age Started School	6.71	1.77	6.64	1.75	-0.07
Minority Language Spoken at Home	0.36	0.48	0.41	0.49	0.05***
Number of Older Siblings	2.44	1.87	2.41	1.84	-0.03
Number of Younger Siblings	1.67	1.48	1.72	1.49	0.04
Both Parents Alive	0.76	0.43	0.78	0.42	0.02
Mother Literate	0.52	0.50	0.49	0.50	-0.03*
Father Literate	0.57	0.49	0.57	0.49	-0.00
Live with Biological Mother	0.73	0.44	0.77	0.42	0.03**
Live with Father	0.55	0.50	0.60	0.49	0.05***
Class Level Variables					
Start-Year Enrolled Class Size	60.08	15.58	60.32	15.90	0.24
Grade Level	4.53	0.50	4.55	0.50	0.01
Observations	2597		2480		5077

Notes: The outcomes end-year math and language test scores have been standardized to a mean of 0 and a standard deviation of 1 in the selected sample. Last column reports the difference between boys and girls with the significance level of the t-test. * p < 0.1, ** p < 0.05, *** p < 0.01.

parents alive, parental literacy, and whether one lives with their biological mother and lives with their father.

In a small number of cases, some student characteristics are missing, as indicated by the count column which reports the number of non-missing observations. For the analysis, we impute these and control for a missing indicator.[22]

[22]We impute age and age at school start to the median if missing. The remaining variables with missing observations are imputed to zero. Similarly, a very small number is missing their start year test scores in which case we impute these to the mean and control for the missing indicators.

We also include two class-level controls. Class size in Ethiopia is often large (Aurino et al., 2014) and in our sample the average size is 60 students, thus we always control for class size. We also control for grade level fixed effects, with the sample nearly evenly split between 4th and 5th grade classes. We do not include region fixed effects as the data we use only gives us the locations of schools, however, once school fixed effects are included region fixed effects would not be separately identified.

4 Empirical Strategy

4.1 Model

We aim to assess the causal effects of a potentially salient feature of school environments: the share of female classmates. In our baseline results, we focus on three important outcomes for the production of human capital collected near the end of the school year ($t + 1$): (i) days absent from school, (ii) math test scores, and (iii) language test scores. While absence from school may indeed impact performance, and thus be a mechanism, given the environment and context we expect that absence from school is important in its own right. Thus, our baseline objective is to estimate the causal effect of the peer composition treatment on each outcome. Our treatment of interest ($\overline{female}_{-icst}$) is the mean (percentage) of female peers in class (c) and school (s) at the start of the school year (t), omitting the individual (i) from the calculation (leave-one-out).

We use the following specification as our preferred model for each outcome:

$$Y_{icst+1} = \overline{female}_{-icst}\beta + W_i'\gamma + X_c'\delta + \eta_s + \epsilon_{icst}, \tag{1}$$

where Y_{icst+1} is one of the baseline outcomes observed at the end of the year; W_i is a vector of child-level characteristics, start of year test scores in both math and language, and a range of additional background characteristics described in Section 3.3; X_c is a vector of classroom level

controls; η_s are school fixed effects; and ϵ_{icst} is the error term. For the test score outcomes, we estimate the model with a standard linear regression.[23] For days absent from school, we use the same specification but account for its count data nature with a negative binomial regression.

Our identification of the causal effect rests on the random assignment of students to classrooms. Thus, we focus on the sub-sample of students randomly assigned.[24] We include a wide range of additional individual controls to enhance precision. We also account for grade-level fixed effects and the student's class size, as classes can be large in Ethiopia, which is true in our data. Furthermore, the school represents the level at which classmate peers can be drawn, thus we remove common shocks at the school level through the inclusion of school fixed effects.

Even with random assignment, it may be that the share of female peers captures other dimensions of peer influence. We have a reduced form specification, thus we do not specifically aim to map each channel through which the share of female peers can work. However, we also consider a range of more restrictive specifications, including the addition of a full set of peer leave-one-out means in start year test scores (math and language) and for each of the individual characteristics. These are reported in the Appendix as part of our robustness checks and return results highly consistent with our baseline.

We focus on estimating the effects separately by gender. Specifically, a number of mechanisms we discuss in the introduction, such as the presence of gendered norms, suggest that the effects may be more important for girls. Our hypothesis is that girls benefit from exposure to more girls in the class potentially through improving beliefs and attenuating the effects of gender stereotypes or through reductions in harmful social interactions such as bullying. Thus, we split the model by gender at the baseline, while we also report results for the full sample.[25]

[23] We also use a linear regression with the mechanisms discussed in Section 5.4.

[24] More precisely we choose the sample of students that are in classrooms listing random assignment as the allocation method and that then pass the Fisher test for balanced assignment of gender across classrooms within the school.

[25] We additionally explore heterogeneities along a range of interesting dimensions. For these we maintain the gender split and then include an interaction between the variable of interest and the peer treatment to maintain statistical power.

4.2 Balance Checks

Random assignment to classrooms, or at least students being as good as randomly assigned, is critical to our identification assumption, as it should eliminate factors that would create selection bias. We now turn to a series of balance checks where we (i) regress a female indicator on the share of female classmates, (ii) assess traditional balance tests on a range of individual characteristics and additionally a set of teacher characteristics, (iii) simulate random re-shuffling of class assignments within schools re-drawing each balance test, and (iv) assess the joint relevance of school by class fixed effects on the share of female peers after removing variation due to school fixed factors.

Effects on gender from the share of female classmates. Under random assignment, there should be no sorting by gender, thus the share of female classmates should not predict own-gender. Similar to Guryan et al. (2009), Getik and Meier (2020), Golsteyn et al. (2020) we assess the association between the own- and peer-level treatment by regressing gender on the share of female peers across four specifications. We begin with school fixed effects and then add further controls. The estimates are reported in Appendix Table A.2. In all specifications, we control for the school level leave-one-out share of female peers, following Guryan et al. (2009), to account for mechanical exclusion bias.[26] Consistent with our expectations we find no statistically significant effect.

Balance checks on additional student and teacher characteristics. In Figure 1, we report point estimates and confidence intervals for each balance test on individual characteristics – including start year test scores as a measure of ex-ante ability – in panel (a) and teacher characteristics in panel (b). In each test, we regress – in separate regressions – the share of female classmates on

[26]Exclusion bias can be induced when regressing own- and peer-level measures. This results because an individual cannot be their own peer, thus if an observation is female, peers in the school who can be drawn always have a lower probability of being female or a higher probability if an observation is male. Caeyers and Fafchamps (2020) show this type of bias is always downward.

the characteristic variable.[27] We control for school fixed effects, a missing indicator for imputed observations where necessary, and in the teacher characteristics an indicator for the math and language teacher being the same person (13% of observations).[28]

Across our balance tests we find generally null results. No individual characteristic is significantly related to the treatment. For teacher characteristics, only one returns a significant estimate (p-value < .05), thus out of 20 tests only 1 fails, which is consistent with random chance. Thus, we continue to find supportive evidence for the random assignment of students to classrooms.

A minor concern is that our minority language indicator may not fully capture ethno-linguistic (and ethno-religious) differences relevant in the Ethiopian context where language, ethnicity, and religious affiliation are highly correlated (Ado et al., 2021). If there are differences between ethno-linguistic or ethno-religious groups in terms of their (gendered) schooling preferences, these might affect selection of boys or girls into schools and correlate with our outcomes. We expect this to be accounted for at the school-region level, thus our school fixed effects will remove these differences, leaving the assignment into classrooms within school uncontaminated. Nevertheless, our balancing test in Figure A.1 shows that no single minority language spoken at home predicts the share of female peers variable, and in our later robustness checks, we include a full set of language fixed effects and find no sensitivity in our results.

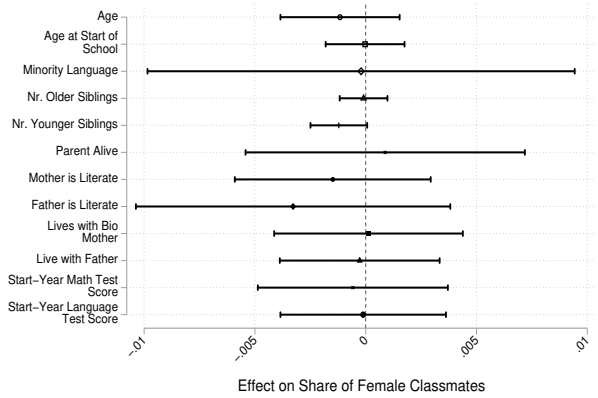
Simulations and balance tests. We next compare the p-values from the balance tests on student and teacher characteristics with those we obtain from randomly re-shuffling students to classrooms within schools. We draw pseudo-random class allocations within school 500 times. At each draw,

[27] Estimating separate regressions means that our balance tests are more conservative compared to estimating a single regression with multiple right-hand-side variables. Nonetheless, we checked whether multiple regressions of all student and teacher characteristics in Figure 1 lead to similar results. The multiple regressions for student and teacher characteristics give F-test of joint significance p-values of 0.1693 and 0.1734, respectively, suggesting that there is no evidence that our covariates are jointly significant in predicting variation in the treatment variable.

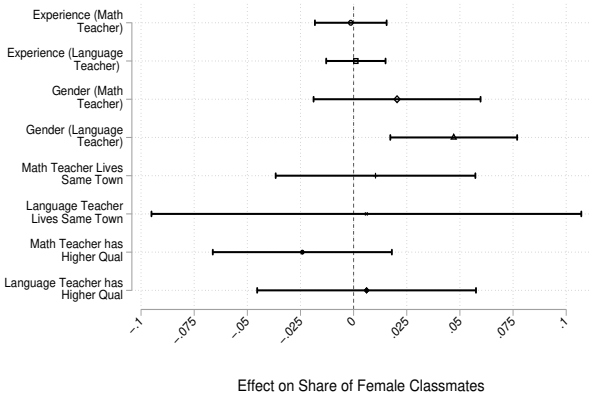
[28] To maintain our full sample, where an observation is missing the characteristic, we impute it to the mean – or zero if an indicator – and control for a missing indicator. Additionally, for teacher experience, we standardize the variables to mean zero and standard deviation of one because the confidence intervals were quite small and seeing their scale is easier with the normalization.

Figure 1. Balancing Tests on Characteristics

(a) Student Characteristics



(b) Teacher Characteristics



Notes: N=5077 in all cases. We regress the share of female peers on each variable (in separate regressions) on the vertical axis. In panel (a) the right hand side variables are student characteristics plus their start year test scores as a measure of ex-ante ability. In panel (b) the right hand side variables are teacher characteristics. The whiskers indicate 95% confidence intervals.

we obtain the placebo share of females from the reallocated class and re-run each balance test conditional on school fixed effects. We then calculate the empirical cumulative distributions for the p-values on the balance tests given the actual class allocations and the pseudo allocations. In comparison, if the actual assignments are random, then we would expect the frequency they are significant to be no greater than the pseudo allocations.[29]

Panel (a) of Appendix Figure A.4 reports these comparisons. We report the means of the empirical CDF for the simulated p-values from 20 equally spaced bins and also the scatter plot of the empirical CDF for the actual values.[30] We find that the actual reject rates at traditional significance levels are very similar to those obtained from the pseudo allocations. We observe no more rejections than would be expected with random noise. We then repeat this comparison in panel (b) using the sample of students who are not randomly assigned. Here we find a higher frequency of reject rates at lower p-values than would be expected from the simulations consistent with concerns

[29]This strategy is similar to that found in Chetty et al. (2009) and Huang et al. (2021).
[30]We use 20 bins because we have 20 individual balance tests in total.

over sorting into classrooms in this non-randomly assigned sample. Thus, the balance test results on our selected, random assignment sample are highly consistent with the random allocation of students to classrooms.

Gender and class fixed effects. Finally, after removing variation due to school fixed effects – the level assignment – we assess whether school by class fixed effects are jointly significant in predicting gender. This follows Chetty et al. (2011), Balestra et al. (2020), Getik and Meier (2020) and supposes that given random assignment school by class fixed effects should not represent relevant predictors of observable characteristics after accounting for school fixed effects.[31] Because class gender composition is our focus, we focus on gender for this test. We first obtain the residuals from regressing an indicator for being female on school fixed effects, and second, we regress these residuals on the school by class fixed effects. We also repeat this adding our baseline set of controls to the first step. In both cases, we find jointly insignificant school by class fixed effects ($F = 0.66$ and $F = 0.64$).

4.3 Additional Concerns

A particularly salient concern for the identification of peer effects is measurement error. Where assignment is not random its bias can be non-classical, resulting in an overestimation of the peer effect, because positive selection on the variable that constructs the peer treatment implies the inclusion of two positively correlated mismeasured regressors (Angrist, 2014, Feld and Zölitz, 2017). The omitted measurement error for the peer variable then contains this positive correlation leading to upward bias. However, with random assignment this correlation has been severed and Feld and Zölitz (2017) demonstrate that in this case measurement error reverts to classical attenuation bias.

[31] We further would not expect a relationship given we remove observations failing the Fisher test that is conducted school by school and tests for balance of gender across classrooms.

We use classrooms that are allocated through random assignment and our balance tests provide strong evidence consistent with random assignment. Moreover, we do not expect that the share of female classmates is measured with substantial error. The Young Live Survey in Ethiopia interviewed everyone in the school who were in grades 4 and 5 and present at the start of school year survey collection. Comparing the number of students we observe in each class to the enrollment number from the class roster, at the start of the year we on average observe 98% of the enrolled class size. Thus, measurement error is not a salient issue in our case, and to the extent there is measurement error, based on random assignment our estimates will be attenuated.

Simultaneity bias is another threat common in the peer effects literature (Sacerdote, 2014). However, we (i) use a pre-determined peer characteristic for our peer treatment and controls, and we (ii) estimate reduced form specifications rather than focus on the peer effect of the outcome. Finally, in our robustness checks where we include peer test scores as controls, we use the start-year measure to minimize the presence of simultaneity bias in the model.

We next present results for the baseline and then key heterogeneities with a focus on gender. We then discuss a number of robustness checks to (i) account for possible nonlinearities in peer start-year test scores, (ii) evaluate additional specifications with higher dimensional controls, (iii) and to assess sensitivity to unobservable selection.

5 Results

We now turn to the results and begin with a set of baseline effects from the share of female classmates on important outcomes for educational development: days absent from school and math and language test scores.[32] Given that the potential mechanisms we discussed can suggest either symmetry or asymmetry in the effect across gender, we begin in Section 5.1 by examining the effect at the mean for females and males. In Section 5.2 we assess a range of robustness checks and then

[32]We use classmates and peers interchangeably.

turn in Section 5.3 to consider a set of heterogeneities. In Section 5.4, we explore for evidence around potential mechanisms, and finally, in Section 5.5, we test for a moderating role from child work.

5.1 Baseline Outcomes

In Table 2, we present the results for the effect of the percentage of females in the class on our baseline set of outcomes. Standard errors are always clustered at the school level. Panel A contains the coefficient estimates based on a negative binomial regression for days absent and linear regressions for standardized test scores. We always include our preferred controls as defined in Sections 3.3 and 4.1 and estimate the models separately by gender.[33]

We find that for females, but not males, an increase in the share of female classmates significantly reduces the number of days absent from school and improves math test scores, while having no effect on language scores. For days absent from school, the average marginal effect among females based on the negative binomial regression is approximately 10.5 fewer days of missed school over the year for a shift from 0% to 100% of the share of female peers.[34] Put in terms of a standardized shift, Panel B shows that a standard deviation shift (9 percentage points) of female classmates translates into a marginal effect of about one less missed day of school (0.95) or 18% of the mean of days absent. For females and math test scores, a standard deviation shift in the share of female peers translates into approximately a 7% of a standard deviation gain.[35]

Our results are asymmetric. Among males, we find no effects. Moreover, we find that coefficients across female and male estimates for both days absent and math test scores are statistically different,

[33] Our results are largely robust to excluding all controls from our baseline regressions, but our estimates do become less precise (see Appendix Table A.3). In robustness checks below, we consider a wide range of additional controls and apply the pdslasso procedure to the selection of control variables. Our results remain robust to the choice of controls included.

[34] We show in the table that the marginal effect on days absent based on an ordinary least squares regression is similar to what we find with the negative binomial but less efficient.

[35] A larger shift in the share of female classmates from the 10th to 90th percentile (23% shift) translates into about 17% of a standard deviation shift in math test scores and about 2.4 fewer days of missed school.

rejecting the null of equality. These results strongly suggest that in our sample males do not and females do benefit from exposure to more female classmates. This is in-line with a number of mechanisms that could generate asymmetric effects such as differences in beliefs driven by gender stereotypes, a reduction in bullying towards girls, or through more attention from teachers.

Between symmetry and asymmetry the current empirical literature, based on data in developed countries, finds mixed results on educational attainment (see, among others, Hoxby, 2000, Lavy and Schlosser, 2011, Black et al., 2013, Cools et al., 2019, Mouganie and Wang, 2020).[36] While in a non-Western context, evidence from middle schools in China indicates that boys benefit more than girls from being exposed to female peers (Gong et al., 2021). Overall, our results where females drive any effects from classmate gender composition seem consistent with a context that is characterised by stronger gendered norms, as opposed to richer contexts studied by previous literature.

To put our findings and their magnitude into the context of similar studies, we consider Gong et al. (2021), the closest paper examining the effect of female peers on test scores among middle school graders in China.[37] Our effects on girls are quantitatively similar to theirs, however, on the contrary to our paper, the latter finds higher effects for boys. As large part of their effects is driven by change in teacher behavior in terms of time allocation and interaction with students, student effort, and classroom environment, this might suggest a more limited role for gender stereotypes or reduction in bullying towards girls.

Our effects are bigger than what has been found for richer countries: Lavy and Schlosser (2011) is the closest paper that examines the effect of female peers on math test scores among 5^{th} graders in Israel, though they consider peers at the level of school grade as opposed to classroom level peers

[36] It is worth nothing that many of these studies finding asymmetric effects focus on peers during the adolescent period rather than the grade range we observe. However, students in our data are on average between 11-12 years old on the cusp of transitioning from childhood to adolescence.

[37] Duflo et al. (2011) and Hahn et al. (2019) analyses peer effects in a developing country context as well, however their papers does not examine peer gender directly.

as in our case.[38] Our larger effects suggest that the role of peer gender can be more influential in a developing country environment than in a richer context such as Israel. This might be driven by several factors: class sizes are much smaller in Israel,[39] where at the same time schools have more resources such as higher teacher/pupils ratio, and teachers are likely to be exposed to more incentives. All such factors point to a more prominent role of teachers as opposed to peer effects in a school environment that is richer than the one we study. Nonetheless, it is also possible that, quite simply, peer gender effects differ across these two contexts due to the fact that stronger gender norms make the effects of peer gender particularly salient in the Ethiopian context.

In the Appendix Table A.4, we report the mean effects in the full selected sample. These are much smaller and not significant, as expected given the strong asymmetry across gender. Thus, going forward we maintain the gender split.

In panel C, we implement some hypothesis testing adjustments and sensitivity diagnostics. One, we cluster the standard errors on schools but we have 41 schools, which for clustering is borderline a safe number of groups.[40] Therefore, we check our results calculating the *p-values* for the test on the coefficients for the share of female classmates based on the Wild cluster bootstrap, which can perform better than standard clustering when the number of clusters is small (Cameron et al., 2008, Roodman et al., 2019). In all cases, inference is unchanged.

Second, we are testing multiple hypotheses. Thus, using the simulated *t-values* from the Wild cluster bootstrap, we implement the Romano-Wolf (RW) multiple hypothesis testing adjustment to control for the family-wise error rate (Romano and Wolf, 2005).[41] We only record small increases in the *p-values* and maintain a 5% significance level for both days absent and math test scores.

[38] In Lavy and Schlosser (2011), a 9 percentage point shift of female school-grade peers translates into a 3.2% of a standard deviation gain for 5^{th} grade girls as opposed to approximately a 7% of a standard deviation gain for our sample of 4^{th} and 5^{th} graders at the classroom level.

[39] Average class size is 60 in our study, whereas in Israel maximum class size is capped at 40 (Angrist and Lavy, 1999).

[40] Cameron and Miller (2015) note that for clustering there is not a clear definition of "few" in terms of the number of groups. It can depend on the situation.

[41] Specifically, we implement the efficient algorithm RW described in Romano and Wolf (2016). To implement this with the Wild cluster bootstrap, we developed a Stata program, *wildrw*, and have made this available at `https://jonathan-norris.github.io/addmat/`.

Table 2. Baseline Outcomes and the Share of Female Peers

	Days Absent from School		Math Test Scores		Language Test Scores	
	(1) Female	(2) Male	(3) Female	(4) Male	(5) Female	(6) Male
Panel A: Baseline Estimates						
Share Female Classmates	-1.99**	-0.64	0.75***	-0.28	0.02	-0.13
	(0.79)	(0.80)	(0.25)	(0.31)	(0.22)	(0.21)
Own-Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Start-Year Test Scores	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2597	2480	2597	2480	2597	2480
R^2			0.605	0.624	0.717	0.733
Equality of Coefs. (p-value)	0.023		0.000		0.469	
ME (nbreg)	-10.53**	-3.70				
	(4.26)	(4.68)				
OLS ME (Days Absent)	-8.64*	-4.58				
	(4.28)	(6.19)				
D.V. Mean by Gender	5.25	5.79	-0.02	0.02	0.07	-0.08
D.V. SD by Gender	(6.71)	(7.89)	(0.98)	(1.02)	(0.99)	(1.00)
Panel B: Standardized Marginal Effects						
Share Female Classmates	-0.95**	-0.33	0.07***	-0.03	0.00	-0.01
	(0.38)	(0.42)	(0.02)	(0.03)	(0.02)	(0.02)
Panel C: Inference and Sensitivity Testing						
Wild Cluster *p-value*	0.031	0.461	0.013	0.402	0.936	0.565
RW *p-value*	0.038	0.502	0.016	0.442	0.936	0.591
Oster's δ ($R^2_{max} = 1.3R^2$)	2.13	1.01	2.19	-0.38	0.02	-0.06

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. All specification are estimated on the sub-sample indicating students were randomly assigned to class and that pass the Fisher test for balanced assignment of gender across classrooms in each school. Columns (1) and (2) are estimated by a negative binomial regression. The treatment variable is the share of female classmates, which is the leave-one-out mean of female peers to the individual in the classroom. In Columns (1) and (2) the dependent variable is the number of days a student is absent from school over the schoolyear. In Columns (3) and (4), the dependent variable is end of the year math test scores for each student. Columns (5) and (6) show the same results for end of the year language test scores. End of year test scores are standardized. All specifications include controls for class size, an indicator for if the class was taught continuously together over the academic year, and school fixed effects. For own-characteristics, we include a gender indicator, a quadratic in age and also in age started school, a home language minority indicator, a control for the number of older siblings and one for the number of younger siblings, an indicator for whether at least one parent is alive, indicators for whether the mother and father are literate, an indicator for whether they live with their biological mother, and an indicator for presence of the father in the home. Where the own characteristics are missing, we impute these (to the mean if continuous, to the median for age, and to 0 if in levels) and control for a missing indicator. In panel B, we report effects on the standardized share of female peers, and in place of the negative binomial coefficient, report the marginal effect (ME) based on this standardization. In panel C, we report *p-values* from the Wild cluster bootstrap and also the Romano Wolf (RW) adjustment for multiple hypothesis testing based on the Wild cluster. Oster's delta is calculated with a $R_{\max} = 1.3 * R^2$ as suggested by Oster (2019). In columns (1) and (2) Oster's delta is calculated from an OLS regression corresponding to the same specification as the negative binomial.

Third, we adopt a more formal approach to sensitivity testing developed in Oster (2019) and calculate the degree of selection on unobservables relative to the selection on our observables (δ)

that would eliminate our observed effects.[42] Values of δ larger than one imply that for the effect to be wiped out selection on unobservables must be larger than selection based on our observables. Where our effects are significant, we would expect values of δ to be at least one given our identification strategy. Indeed, among females we find δ values above two for both days absent and math test scores. These results strongly suggest that our results are not sensitive to unobservables.

5.2 Robustness Checks

We now turn to a series of robustness checks to test our results against sensitivity. Throughout these checks we continue to estimate the specifications separately by gender.

Nonlinearities in peer start-year skills. In our baseline specification we do not include additional peer means for start-year peer test scores. Here we add these. Further, one may be concerned that the share of female classmates captures something about nonlinearities in ability peer effects. Thus, in Appendix Table A.5 we add, in successive regressions, polynomials in math and language test scores from degree one up to four. For each outcome, we find stable estimates for the share of female peers across gender that remain significant for females on days absent and math test scores.

Additional specifications and high dimensional controls. Note, that while our results are largely robust to the exclusion of all controls from our baseline regressions, our estimates do become less precise if we omit start-year test scores (see Appendix Table A.3). We believe that including at least start-year test scores as controls leads to efficiency improvements due to , for example, being able to capture variation in pre-existing differences in skills across students. Nonetheless, this can also lead to concerns that our results only hold up when our specific set of controls is included (specification searching). To allay these concerns, we check whether our results are sensitive to the choice of control variables included in our baseline specifications.

[42] This also requires an assumption about the maximum degree of R^2 that can be allowed. We follow the suggestion by Oster (2019) and use a default $R_{\max} = 1.3 * R^2$.

We begin by expanding the controls over potentially relevant dimensions. In Table 3, we first add a full set of peer means on start-year test scores and for each characteristic in our control set. Second, we include a set of teacher characteristics, as defined in Figure 1. Third, there are a large number of languages spoken in Ethiopia that also may capture ethno-linguistic differences in schooling preferences which may affect boys and girls differently. To control for this, we replace the teacher controls with home language fixed effects. Finally, we add to our main control set the full set of additional peer means, teacher characteristics, home language fixed effects, and through a 5th degree polynomial in start-year tests scores, peer start-year test scores, and all additional peer characteristics. In each iteration of additional controls, we find our treatment effect to remain stable and with a similar significance.

Table 3. Additional Specifications and High Dimensional Controls

	OLS (unpenalized)						PDS Lasso	
	(1) Female	(2) Male	(3) Female	(4) Male	(5) Female	(6) Male	(7) Female	(8) Male
Panel A: Days Absent								
Share Female Classmates	-10.00**	-5.27	-9.19**	-3.92	-10.30**	-5.46	-9.02**	-5.09
	(4.23)	(5.78)	(4.25)	(5.38)	(4.32)	(5.91)	(4.01)	(5.08)
Observations	2597	2480	2597	2480	2597	2480	2597	2480
# Unpenalized Controls	37	37	46	46	47	47	0	0
# Penalized Controls	0	0	0	0	0	0	115	115
# Selected Controls	37	37	46	46	47	47	3	2
Panel B: Math IRT Scores								
Share Female Classmates	0.58**	-0.46*	0.56**	-0.45*	0.57**	-0.48*	0.70***	-0.33
	(0.24)	(0.26)	(0.27)	(0.24)	(0.24)	(0.25)	(0.26)	(0.29)
Observations	2597	2480	2597	2480	2597	2480	2597	2480
# Unpenalized Controls	37	37	46	46	47	47	0	0
# Penalized Controls	0	0	0	0	0	0	115	115
# Selected Controls	37	37	46	46	47	47	2	3
Panel C: Language IRT Scores								
Share Female Classmates	-0.11	-0.32	-0.19	-0.36	-0.11	-0.33	0.05	-0.07
	(0.25)	(0.23)	(0.23)	(0.25)	(0.24)	(0.23)	(0.20)	(0.22)
Observations	2597	2480	2597	2480	2597	2480	2597	2480
# Unpenalized Controls	37	37	46	46	47	47	0	0
# Penalized Controls	0	0	0	0	0	0	115	115
# Selected Controls	37	37	46	46	47	47	2	3
Add Peer Means	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Add Teacher Characteristics	No	No	Yes	Yes	No	No	Yes	Yes
Add Home Language FEs	No	No	No	No	Yes	Yes	Yes	Yes

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are in parentheses and clustered at the school level. Panel A summarises results for the days absent outcome variable. Panel B summarises results for end of year math scores, while Panel C shows results for end of year language scores. Columns 1 and 2 report OLS estimates adding a full set of peer means for start-year test scores and student characteristics as controls. Columns 3 and 4 adds teacher characteristics (missing teacher information is imputed and a missing indicator controlled where necessary) plus an indicator for whether the math and language are taught by the same person (only 13% of the data). Columns 5 and 6 replace the teacher controls with a full set of home language fixed effects. Finally, columns 7 and 8 report estimates after the post-double selection (PDS) Lasso method developed in Belloni et al. (2014) using the theory driven penalizer developed in Belloni et al. (2012). Specifications for the PDS Lasso include all baseline, teacher, and home language fixed effect controls and add through a 5th degree polynomial in start-year test scores and all peer controls. All specifications include, and do not penalize school fixed effects, as even with random assignment accounting for common shocks at the level of student sorting is important. The key peer treatment variable is not penalized and inference on it is valid. Counts of the number of included unpenalized and penalized controls do not include the school fixed effects – there are 41 schools.

Next, we include all of our original controls – including start-year test scores – and additional controls just listed into a single specification. This set, not including school fixed effects, contains 115 controls. We then use a post-double selection (PDS) lasso (Belloni et al., 2014) to select the controls that are the best predictors of both the outcome and peer female composition and include the union of selected controls from each as the control set.[43] Note, that we do not penalize school fixed effects because accounting for shocks at the level of random assignment, the school, is still important. Inference is not valid on the selected controls, however, Belloni et al. (2014) show that it remains valid for the treatment, in our case the share of female peers. The evidence is again highly consistent with our baseline results. Among females the PDS lasso only selects 3 controls for days absent and 2 for test scores (2 and 3 among males) and returns effect sizes on the share of female classmates that are very similar in magnitude and significance to our baseline.

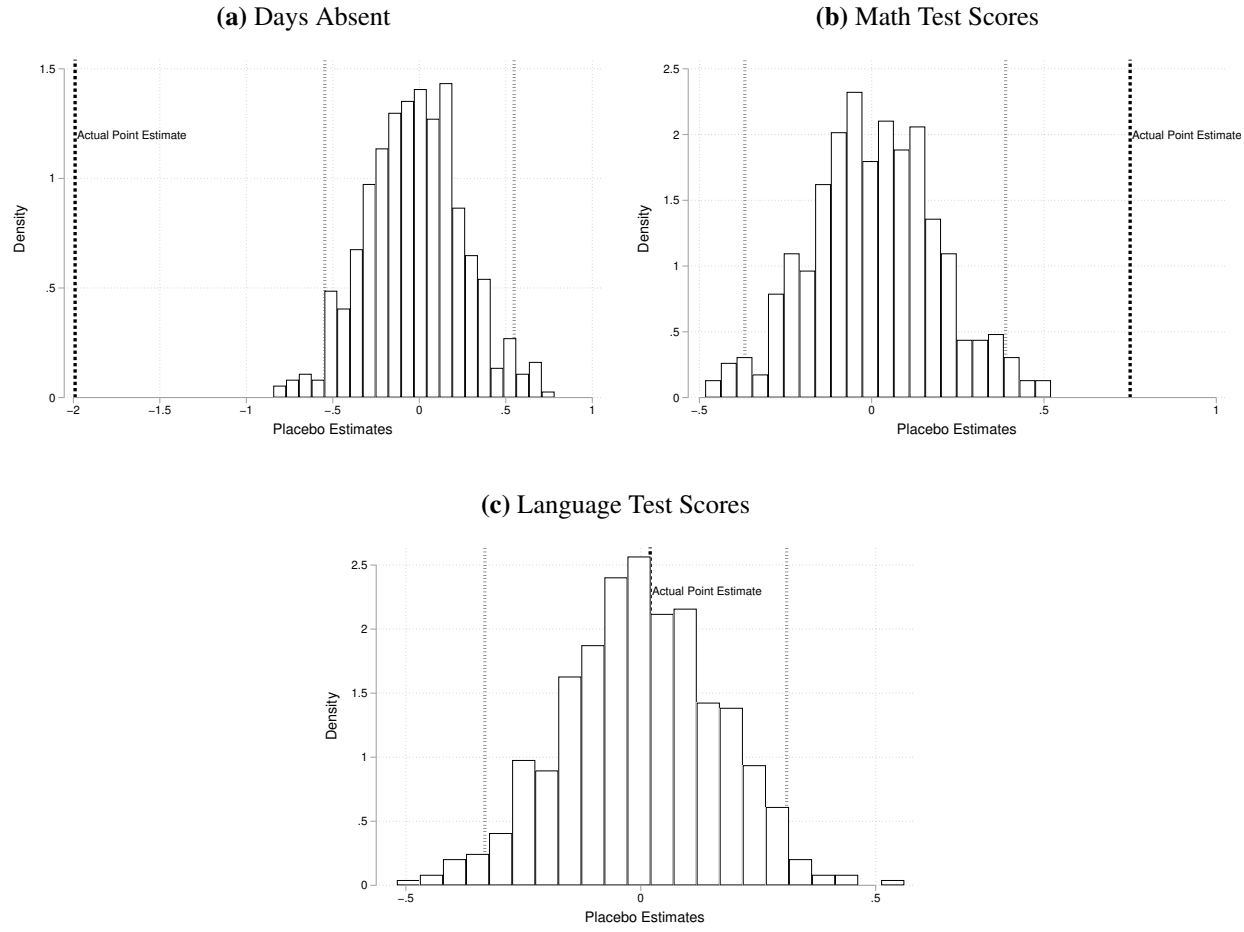
Unobservables and selection: placebo tests. For identification, we leverage the teacher report that students were randomly assigned to the classroom. The results from our balance checks reported in Section 4.2 are consistent with the assumption that these students are indeed randomly assigned. Further, as discussed in Section 5.1, for our primary results, we calculate Oster's δ as the degree of selection based on unobservables relative to observables required to wipe out our estimated effects (Oster, 2019). Based on this diagnostic, we find our results to be highly robust.

As an additional check, we randomly re-shuffle students within schools to classrooms, re-estimate the effect for each outcome by gender, and repeat this for 500 repetitions. Our expectation is that the estimates based on the true share of female classmates should fall in the far tail of the distribution of simulated estimates. In Figure 2, we report the distribution of effect estimates for females and indeed find this is the case. The simulated effects are approximately normally distributed about 0, and where our actual effects were strong and significant (outcomes: days absent

[43] For simplicity we estimate the days absent model with an OLS but as shown in Table 2 it returns very similar, if less efficient, results.

and math test scores), they fall entirely outside the distribution of simulated effects.[44]

Figure 2. Histograms of Permutation Tests: Random Re-shuffle of Students to Classrooms (Female Sample)



Notes: We randomly re-allocate students within schools to the classrooms, holding the number of classrooms to the number we observe in each school, and then recalculate the peer information and regression estimates. We repeat this over 500 repetitions. The true estimate is marked by the vertical dashed line and labeled, while the vertical dotted lines on the ends of the histogram represent the 2.5 and 97.5 percentile points of the simulated estimates. In panel (a), we show the histogram of the estimate from the negative binomial regression of days absent on our preferred specification from column 1 of Table 2. Panel (b) similarly reports results for math test scores and panel (c) for language test scores.

Taking the combination of our checks together, we conclude that our results are not sensitive and are consistent with our assumption of causal estimates based on the random assignment of

[44]We report our results on males in Figure A.5. Here we find the actual point estimates always fall within the distribution of simulated estimates, as we would expect given the actual point estimates for males are closer to zero and not significant.

students to classrooms.

Attrition. We also examine whether our estimates are robust to attrition. Depending on whether attrition is selective, it can potentially affect the gender composition of the classroom, and therefore our peer gender variable. Following the standard approach in the literature (Baranov et al., 2020, Lubotsky, 2007), we correct our baseline specification using inverse probability weighting (IPW). Table A.6 shows that attrition is negatively correlated to the share of same gender peers (Column 1) for the full sample, and that this correlation is driven by the male sample (Column 3). Next, we correct our baseline estimates reported in Table 2 with the inverse probability weights obtained from the estimates in Table A.6. Our results suggest that, despite female peers reducing attrition for boys only, once we correct our estimates with the estimated IPW, we show that for both girls (Table A.7), and boys (Table A.8), baseline estimates are robust to correcting for attrition.

Under-reporting of school absences. Finally, we examine whether our results on school absences are sensitive to potential non-classical measurement error from systematic under-reporting of school absences. The days absent variable is reported through the class roster, which is filled in by teachers at the end of the school year (see Section 3). In general, the mean number of absences in our sample (5.5) is not particularly high for a developing context, but this is largely consistent with the fact that our sample mostly contains schools located in urban areas, where absenteeism is lower.[45] Moreover, our sample is balanced across observable teacher characteristics, and the inclusion of school fixed effects in our baseline specification ensures that our results are not biased even if features of the school environment lead to systematic under-reporting of absences. Nonetheless, the tendency to underreport by teachers might be unobservable and correlated with classroom gender composition (through gender bias, for example).

To illustrate when our result would be sensitive to this possibility, we conduct simulations with non-classical measurement error through under-reporting. Our approach and the simulation

[45]The mean number of absences in rural schools in our sample is 8.17 days per school-year.

results are summarised in the Appendix Section B. We show that the only way in which bias from under-reporting could change our point estimates is through a very strong link between the variance of under-reporting and the share of female peers. Under-reporting itself will not bias our estimates absent this link. Given our treatment is strongly balanced, conditional on school fixed effects, across student and teacher characteristics we think it highly unlikely for there to be such a link mitigating the concern that under-reporting biases our effect estimates on days absent from school.

5.3 Heterogeneities by Additional Characteristics

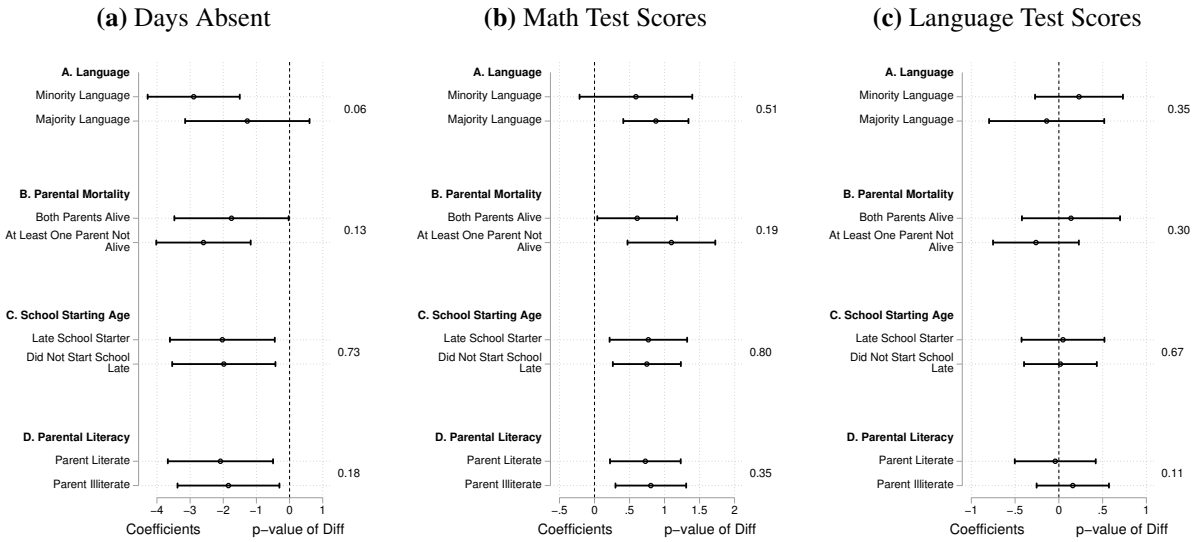
We also assess heterogeneities along characteristics of students that may capture individual disadvantage, of teachers, and at the school or class level. We further explore for non-linearities in the effect of classmate gender composition.

Student characteristics. At the student level, and in separate regressions by gender, we use interactions between the share of female classmates and the following set of indicators: speaking a minority language, parental mortality, late school starters, and parental literacy. The effect estimates and confidence intervals for the share of female classmates by each category of these characteristics are reported in Figure 3, for females, and in the Appendix Figure A.6 for males. Focusing on females, in general the effects tend to be similar across categories.[46] The only exception is on days absent from school and suggests that the effects are larger in magnitude among females who speak a minority language. Minority language speakers might be concentrated in regions, such as Oromia or SNNP, which have considerably lower access to primary education compared to majority language speaking areas such as Addis Ababa (UNICEF Ethiopia, 2019). It is possible that this creates a margin of disadvantage for minority speakers, which is mitigated for females through the effect of sharing the classroom with more female peers. These differences are significant at the 10% level and suggest that the saliency of classmate gender composition may adjust to the external environment.

[46] Among males there are no significant differences between marginal effects across categories for these characteristics.

Nevertheless, we do not find these heterogeneities on math test scores and caution against making strong conclusions based on only this result.

Figure 3. Heterogeneity by Student Characteristics - Female Sample

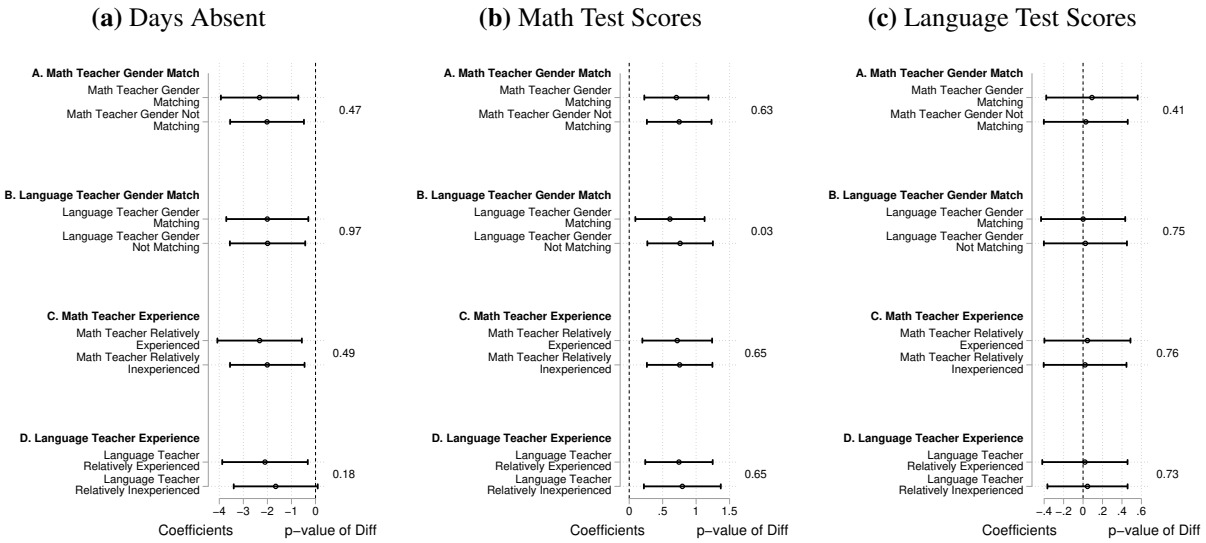


Notes: This figure presents heterogeneous effects of different subgroups on our outcomes including 95% confidence intervals clustered at the school level. We interact the share of female classmates variable with indicators of the respective student characteristics. The dependent variable is days absent in Panel (a); standardised math scores in Panel (b); and standardised language scores in Panel (c).

Teacher characteristics. We report similar heterogeneity results by a set of teacher characteristics for females in Figure 4 and for males in the Appendix Figure A.7. We distinguish between math and language teachers' characteristics, but control for the possibility that the same teacher might teach both subjects in some classes. Again, for both females and males, the effects of the share of female classmates are fairly similar across categories.

School characteristics. In Figure 5, we report heterogeneity by a set of school characteristics for females, whereas the results for males are reported in the Appendix Figure A.8. Our results across categories are mostly similar, although our results for males suggest a negative effect from a higher share of female classmates for those living in rural areas, and in shift schools. In shift schools, one

Figure 4. Heterogeneity by Teacher Characteristics - Female Sample



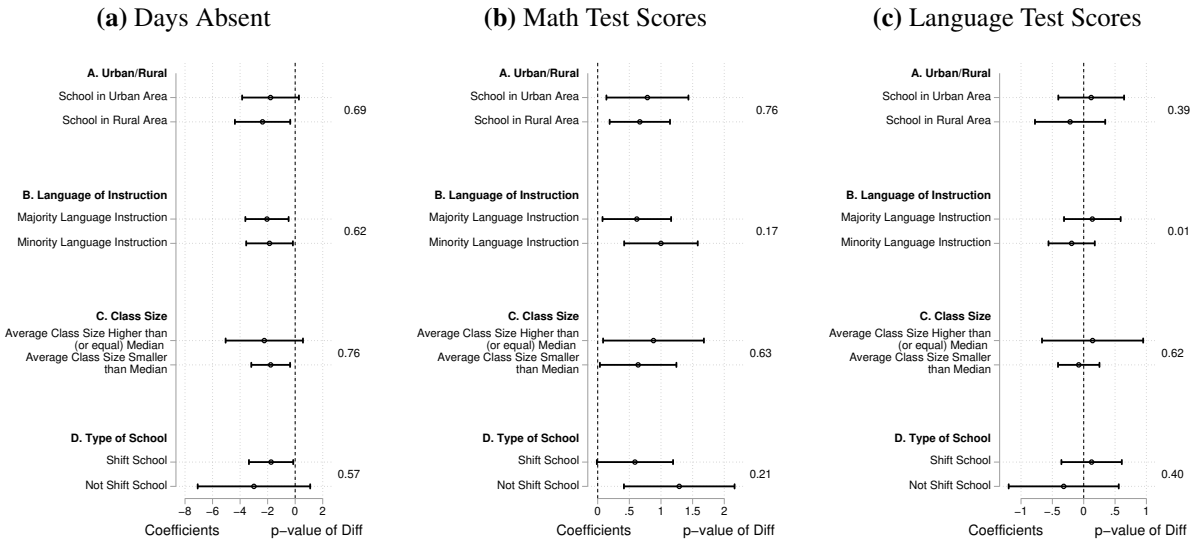
Notes: This figure presents heterogeneous effects of different subgroups on our outcomes including 95% confidence intervals clustered at the school level. We interact the share of female classmates variable with indicators of the respective teacher characteristics. The dependent variable is days absent in Panel (a); standardised math scores in Panel (b); and standardised language scores in Panel (c).

group of students are taught in a morning session, while the other group is taught in the afternoon, while regular (non-shift) schools offer a full day of schooling to students (Orkin, 2013). Possibly, being in a shift school changes the exposure to the peer effect from female classmates, thereby changing the extent to which students could benefit (or detriment) from the peer environment. Nonetheless, for other outcomes and for the female sample, there are no significant differences within this category.

Nonlinear effects. Another feasible dimension of heterogeneity is non-linearity in the effect of classmate gender composition. This would be present, for instance, if the influence of female peers only becomes substantial once their share reaches a critical mass. We check for non-linearity by adjusting our specification from equation (1) to include a quadratic in the share of female classmates.

In Figures A.9 and A.10, we report the marginal effect at deciles of the share of female classmates for females and males. While the quadratic term is not significant, the general pattern does suggest

Figure 5. Heterogeneity by School Characteristics - Female Sample



Notes:This figure presents heterogeneous effects of different subgroups on our outcomes including 95% confidence intervals clustered at the school level. We interact the share of female classmates variable with indicators of the respective school characteristics. The dependent variable is days absent in Panel (a); standardised math scores in Panel (b); and standardised language scores in Panel (c).

some heterogeneity. Among females the effects on days absent and math scores become stronger and significant, as the share of female peers rises beyond the second decile. Among males, we find null effects on test scores across deciles, but on days absent, the effect is negative and significant once the share of female classmates becomes considerably large – beyond the 6th decile. Nevertheless, the results among males remain generally consistent with the asymmetry we find at the mean.

The evidence here is suggestive that the impact of female peers grows as they reach larger proportions of the class composition. This would be consistent with a number of mechanisms. Shifts in the saliency of gendered norms and beliefs, changes in bullying or class behavior, or shifts in teachers' attention may require a sufficient proportion of girls in the class to enable these mechanisms.

The general lack of heterogeneity by student and teacher dimensions at least imply that our gender heterogeneity results do not simply pick up a wide variety of heterogeneities. Rather,

our results point strongly toward effects stemming from female classmates and the presence of a particular asymmetry where effects are focused on females. One limitation of our data is that we do not observe parental or teacher beliefs about ability across gender. This precludes us from assessing heterogeneous effects as a moderating role for this potential mechanism. However, in Section 5.4 we are able to explore a number of channels related to potential mechanisms, which we turn to next.

5.4 Mechanisms

In motivating our focus on peer gender, we discussed some potential mechanisms that fall under either social interactions or shifts in teacher behavior. In this section, we assess factors in our data that can point us toward likely mechanisms and suggest how these effects may work.

5.4.1 Motivation and Participation in Class

In addition to our baseline outcomes, at the end-of-year survey we also observe for each student a ten point motivation scale and another for class participation. These are reported by the teacher. As we discuss in the introduction, experimental results have found females to withdraw more from competition and lower their beliefs on gender stereotyped categories in the presence of males (Bordalo et al., 2019, Niederle et al., 2013). Gender norms are strong in Ethiopia. To the extent that this drives more extensive gender stereotypes, more boys in the classroom could act to lower girls' motivation and participation and represent a direct social interaction effect from peers. Alternatively, if exposure to more females in the class shifts teachers' attitudes or treatment toward girls, then we would again expect positive effects on girls working through a teacher mechanism.

In Table 4, we report estimated effects from the share of female classmates on end-of-year motivation and participation.[47] For each, we report the estimated effect among all students and then split by gender.

[47] Both of these have been standardized to mean zero and a standard deviation of one.

Table 4. Motivation and Class Participation

	Motivation (z-score)			Participation in Class (z-score)		
	(1) All	(2) Female	(3) Male	(4) All	(5) Female	(6) Male
Share Female Classmates	1.62**	1.82**	1.41**	1.25***	1.49***	1.03*
	(0.65)	(0.76)	(0.66)	(0.45)	(0.53)	(0.52)
Own-Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Start-Year Test Scores	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5077	2597	2480	5077	2597	2480
R^2	0.293	0.293	0.316	0.315	0.324	0.326
D.V. Mean	-0.00	0.01	-0.01	0.00	0.01	-0.01
D.V. SD	(1.00)	(0.98)	(1.02)	(1.00)	(1.00)	(1.00)
Equality of Coefs. (p-value)		0.450			0.367	
Oster's δ ($R^2_{max} = 1.3R^2$)	-2.23	-2.69	-1.79	-1.58	-1.84	-1.28

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. The treatment variable is the share of female classmates, which is the leave-one-out mean of female peers to the individual in the classroom. Motivation and class participation are from 10 point scales collected at the end-year survey. These are standardized for the regression. All other definitions and specifications are as defined in Table 2. Oster's delta is calculated with a $R_{\max} = 1.3 * R^2$ as suggested by Oster (2019).

For both outcomes, there are strong effects among females, but unlike the baseline, results are symmetric: females and males improve with more female peers.[48] The point estimates are larger for females but not statistically different from those for males.[49]

These results are consistent with our baseline estimates where we find that effects are located among females, but also, suggest that boys are affected and improve in motivation and participation as the share of females rises in the class. These benefits for boys could stem from a number of channels such as less disruptive class environment, similar to what Lavy and Schlosser (2011) find with Israeli data, or less competitive social interactions if females tend to compete less stringently. However, these positive benefits for boys do not translate into improvements on missed schooling and test scores.

[48]Among females, an approximate 9 percentage point (1SD) shift in the share of female peers, improves motivation by about 17% and participation by about 13% of a standard deviation. Similarly for males it is 12% and 9%.

[49]Further, in all cases we find values of Oster's delta larger in absolute value than one, implying these estimates are not very sensitive to potential omitted variables.

Our evidence on females is at least consistent with mechanisms that generate asymmetries in our baseline schooling outcomes. In the social interaction case, when there are more girls in a classroom, priming of gender norms may be lower and girls may become more motivated and participate more in the class, translating into fewer missed days of school and better performance. Nevertheless, from these, we cannot rule out that the effects stem from shifts in teacher behavior.[50] Thus, we now turn to investigate some observable teacher behaviors and attitudes, which, though not perfect, may provide at least some insights on potential reactions from teachers.

5.4.2 Teacher Behavior

We use three measures of teacher behavior: absences, the teacher changing before the end of the school year, and a scale of teacher motivation, or belief in their capability, to help their pupils learn. We observe these measures for math and language teachers. In Table 5, we regress each of these by each teacher on the share of female classmates and our baseline control set.

At the end-year survey, we observe information on self-reported absences by math and language teachers. [51]

Teacher absences in Ethiopia are a significant problem, especially in rural schools (Abebe and Woldehanna, 2013, Tafere and Pankhurst, 2015, Tafere and Tiumelissan, 2020). While some absences are likely driven by constraints, such as poor conditions or wages, it also may capture commitment and ability to instruct the classroom. In the event that classroom gender composition affects teachers' motivation – e.g., through better behaved students or via their own gendered beliefs – then it could translate into shifts in absences. In columns (1) and (2), we find null results on both the math and language teacher absences, and while the standard errors are large, the point estimates

[50] If the presence of more girls in the classroom shifts teachers' attention to girls or reduces teacher bias in favour of boys, then in both cases we would observe an improvement in girls' performances (Lavy and Sand, 2018, Gong et al., 2021).

[51] In a small share of instances (15.8%), the teacher changed during the year. In this case, the new teacher was asked about their own and the past teacher's absences and we take sum of both to represent teacher absences for students in that class. These are self-reported, thus may be subject to misclassification. We expect then that the regression estimates on the share of females will be unbiased but inefficient.

are relatively small, with the exception of column (2). We then replace the dependent variable with an indicator for whether the math teacher (column (3)) and language teacher (column (4)) changed during the school year. Again, we find null results, with small point estimates – in terms of a standard deviation shift in the share of females in the class – that are insignificant.

Finally, we construct an index of teacher motivation from a set of items answered by the teachers that rate their beliefs on their ability and motivation to help students learn.[52] Summary statistics for the original survey items are summarised in Table A.10 in the Appendix.[53]

To the extent that teachers hold gendered stereotypes themselves or that classroom gender composition changes classroom behavior, then teacher beliefs and motivations may shift in response to the classroom gender composition. It is feasible this could happen rapidly, if teachers have already formed opinions or past experience with different gender compositions in class. Yet, we again find null results on the share of female peers (columns (5) and (6)).

We find no evidence on these teacher behaviors that they respond to the share of females in the class. Though we might not fully capture shifts in teacher motivation as the latter is measured shortly after the school starts, for the end-year measures for teacher absences and whether the teacher changed, we find no effects. In general, teachers in Ethiopia face many other factors and constraints that likely drive their behavior, such as poor compensation or inadequate facilities(Yadete, 2012, Abebe and Woldehanna, 2013). Some of our measures are measured shortly after the start-year survey, and, in addition, unfortunately, we are not able to explore other aspects of teacher behaviors that might be affected by peer gender, so we need to interpret these results with caution. However, we argue that our results here are not inconsistent with peer gender effects acting directly through social interaction mechanisms in a setting of large classrooms and teachers equipped with poor

[52]A principle component factor analysis returns two components explaining more variation than a single variable but the first component captures most of the variation and the rotated loadings indicate a clear pattern of strong loadings on this first component. We extract this first component based on the rotated loadings, standardize it, and use it as our teacher motivation scale for math and language teachers.

[53]These items are collected at the start-year-survey but not again at the end-year survey. While the start-year survey is near the beginning of the school year, it is still after students have been assigned to class and thus our identification strategy remains valid.

Table 5. Share of Female Classmates and Effects on Teacher Behavior

	Teacher Absences		Teacher Change		Teacher Motivation	
	(1) Math	(2) Language	(3) Math	(4) Language	(5) Math	(6) Language
Share Female Classmates	1.98	4.51	0.39	0.35	1.90	-0.55
	(6.43)	(7.31)	(0.42)	(0.45)	(1.31)	(1.08)
Observations	5077	5077	5077	5077	5012	5003
D.V. Mean	4.03	3.32	0.11	0.09	0.00	0.00
D.V. SD	(5.44)	(5.92)	(0.31)	(0.28)	(1.00)	(1.00)

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. All specifications include our baseline set of controls and are estimated on our analytical sample. The treatment variable is the share of female classmates, which is the leave-one-out mean of female peers to the individual in the classroom. The dependent variables in Columns 1-2 are days absent by class for math (1) and language teachers (2) and in Columns 3-4 they are an indicator for whether the math teacher (3) was changed during the year and similar for the language teacher (4). A dummy variable indicating whether the math and language subjects are taught by the same teacher is included in all specifications. In Columns 5-6, we use the standardized predicted score, for math teachers (5) and language teachers (6), from a principle component factor analysis for each teacher on items related to how well the teacher feels they can motivate and help their students. One component adequately summarized the correlation across these items.

resources and incentives.

5.4.3 Heterogeneity by Classmate Age

Next, we address whether the age of classmates varies the peer gender effect. A feasible mechanism for peer gender effects within social interactions is through shifts in girls' beliefs about capabilities. If social interactions drive the effects, then the presence of older boys could exasperate the problem that exposure to more girls reduces. Another form of social interaction effects would stem from the ability to form friendships (homophily), whereby girls are more likely to create friendships with other girls if they are of similar age.

Conversely, where the effects are driven by shifts in teacher behavior, we would expect to see weaker effects whenever classmates of either gender tend to be older. The idea here is that this may force the teacher to split attention in a way that hinders the progress of girls and boys. For example, if more girls in the class implies a better behaved class, teachers may be able to focus

more on instruction; however, when there are older peers in the classroom (of either gender) this may constrain the teacher's instruction as they divide attention across age groups in a manner that would hinder both girls and boys.

The age distribution of peers is a feature of the classroom environment in Ethiopia that is particularly different from environments studied in the previous literature. As we showed and discussed in Section 3.3, students are on average around the correct age for the grades surveyed but there is significant dispersion due to late starters and likely those who, once in school, repeat grades from missed schooling.[54] In Appendix Figures A.11 and A.12, we first check whether our treatment effect varies by a student's own-age and find no evidence it does. Here, we turn our attention to the heterogeneity by classmates age.

To address how classmates' age matters for peer gender effects, we construct indicators for whether the mean of own-gendered peers' age falls in the top tertile, and likewise, an indicator for whether opposite gendered peers' age falls in the top tertile.[55] We then add to our baseline specification interactions with these indicators, first with each indicator seperately (columns 1, 2, 4, 5, 7, 8) and second with a full set of interactions, i.e., a triple interaction, between these indicators and the share of female classmates (columns 3, 6, 9).[56] While we do not include interactions across all tertiles because of sample size limitations, we are able to address whether the peer gender effect varies by exposure to own- and opposite gender classmates who are on average in the top tertile of the age distribution for their gender. In Table 6, we focus on females and report the marginal effect from the share of female classmates in each combination of female and male classmates' top tertile age indicators. Effects on males are reported in the Appendix, Table A.9.

Our results for girls, in Table 6, suggest some important heterogeneities. For days absent, girls benefit, significantly reducing their absences, from exposure to more girls regardless of whether

[54] Mean classmates age ranges from 10 to 13.4.

[55] For females, the mean age of female classmates within the top tertile is 12.4 and when the mean of males' age is in the top tertile, average male age is 12.45 – both range from approximately 12 to 14.

[56] We also control for tertile fixed effects in own- and opposite gender classmates age along with the mean of all classmates age.

girls in the class are older or younger. However, the presence of older boys – where the mean age of boys' is in the top tertile – always weakens the effect of more girls in the class. One explanation is that when boys tend to be older there is more bullying, discouraging girls' attendance such that benefits of more girls in a class are moderated toward zero. Yet, it could also be that older boys simply change the attitudes of teachers and we cannot observe this in our data, thus we caution stronger conclusions.

Table 6. Peer Gender Effects by Peer Age - Female Sample

	Days Absent			Math Scores			Language Scores		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Marginal Effects: Share Female Classmates by									
Female Peers in Bottom Two Age Tertiles	-10.49*			1.05***			0.19		
	(5.24)			(0.28)			(0.28)		
Female Peers in Top Age Tertile	-9.07			-0.11			-0.26		
	(7.55)			(0.40)			(0.37)		
Male Peers in Bottom Two Age Tertiles		-13.53**			0.88***			0.16	
		(5.12)			(0.30)			(0.25)	
Male Peers in Top Age Tertile		-7.94*			0.73**			0.16	
		(4.58)			(0.26)			(0.25)	
Female Peers in Bottom Two Age Tertiles × Male Peers Bottom Two Tertiles			-13.57**			1.12***			0.24
			(5.53)			(0.32)			(0.24)
Female Peers Top Age Tertile × Male Peers in Bottom Two Tertiles			-13.93**			-0.40			-0.14
			(6.44)			(0.47)			(0.31)
Female Peers Bottom Two Age Tertiles × Male Peers Top Tertile			-5.25			1.30***			0.15
			(10.21)			(0.24)			(0.38)
Female Peers Top Age Tertile × Male Peers Top Tertile			-7.37			0.13			-0.36
			(8.48)			(0.40)			(0.41)
Observations	2597	2597	2597	2597	2597	2597	2597	2597	2597

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. Each column is a single regression. We estimate our baseline specification for each outcome adding (in seperate regressions) interactions between our the share of female peers and an indicator for whether the mean of a students' female peers are in the top tercile of the female peers' age distribution and likewise for male peers mean age falling in the top tercile. We then add a triple interaction between the share of female peers and both of these indicators. We calculate the marginal effects for the share of female peers at each combination of the peer top tercile age indicators and report these in the table. We restrict the sample to contain only females.

On math test scores, we find a different pattern. Here girls benefit from exposure to more girls, as long as those girls do not fall in the oldest age group. The age of boys, however, does not vary the effect. Math test scores capture academic performance, which may be more affected by ability to form friendships motivating effort and confidence (e.g., through beliefs). If this pattern was about teacher shifts in practice, then we would also expect to see weaker peer gender effects when boys are older. Nevertheless, we caution strong conclusions and see these results as suggestive.

Rather than social interaction effects these differences in the marginal effect from the share of female peers could be explained by differences across gender in previous academic trajectories. This could drive our results here through leading to adjustments by teachers, or via differences in class behaviors, that seperate by age across gender. We check this in the Appendix Table A.11 and report conditional mean gaps in gender across three measures of previous academic trajectories. These are ever repeated a grade, ever dropped out, and attended pre-school.

Based on the institutional and cultural background we do not expect much difference across gender by grade 5, especially in our sample which mostly features schools in urban areas, where dropping out is perhaps less likely to be an issue (see Section 5.2). In lower grades, we might even expect girls to have a slightly stronger background as families are known to be more likely to keep young girls in school and around the house (Favara, 2017). This is what we observe. There are no differences across gender, with the exception that girls are slightly less likely to have previously dropped out. Given the institutional background through primary school and the evidence here we think that academic trajectories are less likely to explain our results. Nevertheless, we caution that this is an alternative channel to social interaction effects that could give rise to indirect effects from the share of female classmates.

In general, the peer gender effect among females exhibits strong patterns of heterogeneity that point toward social interaction mechanisms. Turning to boys, reported in Appendix Table A.9, we do not find a significant pattern of effects. This is again consistent with the asymmetry observed in the baseline and with mechanisms driving effects around females. Further, our results point to important sources on motivation and participation in class and along the age of classmates, while indicating the absence of effects on our set of observable teacher behaviors.[57] While we cannot conclude that social interactions drive our results, our evidence is consistent with this channel, and regardless, these results suggest that age differences in classrooms have important consequences for

[57]Using the Younger Cohort (YC) of the Young Lives longitudinal survey, we find suggestive evidence that girls face less bullying when they have a higher share of female classmates (evidence available upon request). Though we cannot draw strong conclusion as the sample size is too small and estimates are imprecisely estimated, bullying might be another important mechanism to explore for future work.

the benefits among girls of sharing the classroom with more girls.

5.5 Moderation by Child Work

The presence of child work in Ethiopia has a strong influence on childrens' time use and tends to be a concern in the broader developing context. Children in Ethiopia might be expected to engage in paid or unpaid work for different reasons. They might work to help their families with domestic or farm activities, or they might be required to generate income through paid labour (Tafere and Pankhurst, 2015). Child work is possibly more of an impediment for boys, who often have to finish or interrupt schooling to do paid work, while girls can balance education with domestic work more flexibly (Favara, 2017, Orkin, 2012).

There is already some evidence from developing countries that the presence of child work might offset the positive effects of early educational influences and investments (Bau et al., 2020). This is because early life shocks that increase returns from education also tend to make child work more attractive by increasing the opportunity cost of schooling. In our context, it is possible that the prevalence of child work for some students reduces exposure to their peers. Moreover, social norms related to child work – which may lead to lower beliefs about children's education – could prevent children from realising improvements at school. All of these channels could lead to child work moderating the positive peer effect from a higher share of female classmates.

To check whether this is the case, we interact our classmate gender composition measure with indicators for whether a student is engaged in more than the median hours spent on different types of child work (farm/family work, paid work, domestic work) during a school day. The median hours spent working in our sample is one for the farm work and domestic work variables, and zero for the paid work variable. We also examine effects by interacting with a categorical variable that takes a value of one when a student in engaged in any child work at all. We report the marginal effects corresponding to these categories in Table 7. Our results are disaggregated across gender and across

different types of child work.[58]

It is clear from Table 7 that the peer effect on both school absences and math scores is considerably stronger (although not significantly different) for females who are less involved in child work.[59] It is possible that the presence of child work makes it harder (though not impossible) for girls to benefit from having a higher share of female classmates.

While child work seems to moderate the positive peer effect, it does not fully offset it, as girls in the high child work categories still benefit from having more female peers. It is worth noting however that this is a short term effect, and it is still possible that these early improvements in educational outcomes will not only increase the returns from education but also the returns from child work, incentivising parental investment in the latter (Bau et al., 2020). In the short-run, where parents may not be able to compare childrens' returns from schooling to returns from child work, it seems likely that social interaction effects from a higher share of female classmates help mitigate the negative effects of child work.

[58] In our sample, 36% of children are involved in farm work, 27.2% are involved in paid work, and 48.1% are engaged in domestic work. Naturally, these categories may overlap and children might engage in more than one of these activities. In fact, 85% of student are involved in some type of child work on a given school day.

[59] One exception to this is the paid work measure, where the peer effect on math scores seems slightly higher for females in the high paid work category.

Table 7. Peer Gender Effects by Degree of Child Work

	Days Absent		Math Scores		Language Scores	
	Female	Male	Female	Male	Female	Male
Panel A: Farm/Family Work						
Marginal Effects (Farm Work = High)	-8.44**	-4.99	0.55*	-0.51	0.28	-0.35
	(4.75)	(4.01)	(0.37)	(0.28)	(0.26)	(0.24)
Marginal Effects (Farm Work = Low)	-12.29**	-2.65	0.91***	-0.11	-0.18	0.03
	(4.88)	(4.70)	(0.30)	(0.32)	(0.30)	(0.21)
p-value of Difference	0.19	0.18	0.26	0.23	0.16	0.10
Panel B: Paid Work						
Marginal Effects (Paid Work = High)	-9.14*	-6.59	0.99***	-0.01	0.27	-0.50*
	(4.92)	(4.45)	(0.32)	(0.42)	(0.30)	(0.29)
Marginal Effects (Paid Work = Low)	-11.14**	-2.53	0.67**	-0.39	-0.05	0.01
	(4.23)	(5.12)	(0.27)	(0.32)	(0.26)	(0.22)
p-value of Difference	0.36	0.29	0.31	0.32	0.39	0.04
Panel C: Domestic Work						
Marginal Effects (Domestic Work = High)	-9.09**	-2.71	0.70**	-0.41	0.19	-0.23
	(4.32)	(4.52)	(0.27)	(0.34)	(0.20)	(0.24)
Marginal Effects (Domestic Work = Low)	-12.77**	-5.21	0.85**	-0.13	-0.24	-0.02
	(5.10)	(5.07)	(0.31)	(0.36)	(0.31)	(0.25)
p-value of Difference	0.37	0.33	0.58	0.38	0.08	0.38
Panel D: Any Child Work						
Marginal Effects (Child Work = Any)	-10.27**	-4.06	0.69**	-0.30	0.03	-0.18
	(4.32)	(4.52)	(0.27)	(0.34)	(0.20)	(0.24)
Marginal Effects (Child Work = None)	-13.48*	-1.03	1.23**	-0.10	-0.05	0.20
	(7.10)	(5.84)	(0.54)	(0.47)	(0.46)	(0.28)
p-value of Difference	0.52	0.35	0.24	0.67	0.83	0.23
Observations	2597	2480	2597	2480	2597	2480

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are in parentheses and clustered at the school level. We run our preferred specification using an interaction term between the share of female classmates variable and a) an indicator for whether a student is involved in more hours of farm/family work than the median b) an indicator for whether a student is involved in more hours of paid work than the median c) an indicator for whether a student is involved in more hours of domestic work than the median and d) whether a student is involved in any child work at all. In Columns (1) and (2) the dependent variable is the number of days a student is absent from school over the schoolyear. In Columns (3) and (4), the dependent variable is end of the year math test scores for each student. Columns (5) and (6) show the same results for end of the year language test scores. End of year test scores are standardized. Any child work is defined as more than one hour of family or domestic work per school day, or any paid work. For the farm work and domestic work variables the median level of child work is one hour per school day. For the paid work variable the median is zero. The p-value of difference indicates that the estimated marginal effects for each binary value of the child work indicators are statistically significantly different from each other.

6 Conclusion

We provide, to our knowledge, the first evidence on the role of classroom gender composition in a developing world context. Based on the random assignment of students to classes in Ethiopia, our analysis provides robust evidence that among girls an increase in the share of female classmates leads to fewer school absences and higher math test scores. The effects on school absences and math test scores are sizeable, and suggest that classmate gender composition is an important determinant of girls' educational outcomes in Ethiopia. Further, these effects are strongly asymmetric. Among boys we find no evidence of a significant effect from classmate gender composition on missed schooling and test scores.

We then show that, among a range of factors sorted around direct, social interaction and indirect mechanisms, our results are consistent with direct effects from peers. We begin by showing that having more females in the classroom strongly increases participation and motivation among girls. Though these effects are symmetric across genders, they appear to only translate into improved attendance and test scores for girls.

We then find a set of results consistent with social interaction effects. First, the share of female peers is not linked to what we observe of teacher behaviours and attitudes towards students. Second, for girls, the effects vary differentially between male and female classmates' age. The benefits on missed schooling are invariant to female classmates' age but are weaker in the presence of older boys consistent with protection effects, while on math test scores the benefits are invariant to male classmates' age but are strongest when other girls are not too old consistent with benefits via friendships. Third, in a small subset of the sample with information on being bullied, we find suggestive evidence that girls experience less bullying when exposed to more female peers. However, we caution that this subset is underpowered to detect effects, thus the evidence here is suggestive for future work.

Due to lack of direct information on parental and teacher preferences and gender bias, we are

not able to test to what extent our findings might be moderated by exposure to stereotypical parental or teacher biases reinforcing gender stereotypes. This could be a useful extension for further work in a developing context, as gender bias is found to affect girls performances (Favara, 2017, Alan et al., 2018).

Finally, we turn to investigate whether child work moderates the influence of female peers. Our results suggest that girls who spend more than an hour per day doing child work experience reduced benefits from female classmates, although both groups continue to benefit significantly. Thus, circumstances outside of the school may play some part in moderating how features of the school environment affect students. Nevertheless, we continue to find even girls' engaged in work benefit from more female peers. We think this is an important indicator that peer features of school environments can be important even when the outside of school environment is less conducive to education.

We believe that understanding how features of the school environment affect students in a developing context is a fruitful area for further research. As education policy within Ethiopia begins to boost more children into education, it will be important to understand the role of peers, teachers, and school policies in keeping children in school and building long-term success. This study shows that the class gender composition is particularly important for girls on attendance, math performance, and motivation.

References

W. Abebe and T. Woldehanna. *Teacher training and development in Ethiopia: Improving education quality by developing teacher skills, attitudes and work conditions*. Young Lives, 2013. 8, 38, 39

D. Ado, A. W. Gelagay, and J. B. Johannessen. The languages of ethiopia. *Grammatical and Sociolinguistic Aspects of Ethiopian Languages*, 48:1, 2021. 11, 18

G. Akerlof and R. Kranton. Economics and identity. *The Quarterly Journal of Economics*, 115(3): 715–753, 2000. 3

S. Alan, S. Ertac, and I. Mumcu. Gender stereotypes in the classroom and effects on achievement. *Review of Economics and Statistics*, 100(5):876–890, 2018. 48

J. Angrist. The perils of peer effects. *Labour Economics*, pages 98–108, 2014. 20

J. D. Angrist and V. Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114(2):533–575, 1999. 4, 24

J. D. Angrist, V. Lavy, J. Leder-Luis, and A. Shany. Maimonides' rule redux. *American Economic Review: Insights*, 1(3):309–324, 2019. 4

E. Aurino, Z. James, and C. Rolleston. Young lives ethiopia school survey 2012-13: Data overview report. Technical Report Working Paper 134, Young Lives, 2014. 9, 13, 15

S. Balestra, B. Eugster, and H. Liebert. Peers with special needs. *Review of Economics and Statistics*, 2020. doi: https://doi.org/10.1162/rest_a_00960. 20

V. Baranov, S. Bhalotra, P. Biroli, and J. Maselko. Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial. *American Economic Review*, 110(3):824–59, 2020. 31

N. Bau, M. Rotemberg, M. Shah, and B. Steinberg. Human capital investment in the presence of child labor. Technical report, National Bureau of Economic Research, 2020. 6, 44, 45

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012. 28

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014. 28, 29

M. Bertrand and J. Pan. The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, 5(1):32–64, 2013. 3

J. Bietenbeck. The long-term impacts of low-achieving childhood peers: Evidence from project star. *Journal of the European Economic Association*, 18(1):392–426, 2020. 5

R. Bifulco, J. M. Fletcher, and S. L. Ross. The effect of classmate characteristics on post-secondary outcomes: Evidence from the add health. *American Economic Journal: Economic Policy*, 3(1): 25–53, 2011. 5

R. Bifulco, J. M. Fletcher, S. J. Oh, and S. L. Ross. Do high school peers have persistent effects on college attainment and other life outcomes? *Labour Economics*, 29:83 – 90, 2014. 5

S. E. Black, P. J. Devereux, and K. G. Salvanes. Under pressure? the effect of peers on outcomes of young adults. *Journal of Labor Economics*, 31(1):119–153, 2013. 2, 23

A. Booth and P. Nolen. Choosing to compete: How different are girls and boys? *Journal of Economic Behavior & Organization*, 81(2):542–555, 2012. 3

P. Bordalo, K. Coffman, N. Gennaioli, and A. Shleifer. Beliefs about gender. *American Economic Review*, 109(3):739–73, 2019. 3, 36

J. Boyden, C. Porter, and I. Zharkevich. Balancing school and work with new opportunities: changes in children's gendered time use in ethiopia (2006–2013). *Children's Geographies*, pages 1–14, 2020. 2

S. Brown and K. Taylor. Bullying, education and earnings: Evidence from the national child development study. *Economics of Education Review*, 27(4):387 – 401, 2008. 4

B. Caeyers and M. Fafchamps. Exclusion bias in the estimation of peer effects. Technical Report DP14386, CEPR Discussion Paper Series, 2020. 17, 60

C. Cameron and D. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. 24

C. Cameron, J. Gelbach, and D. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008. 24

S. E. Carrell, M. Hoekstra, and E. Kuka. The long-run effects of disruptive peers. *American Economic Review*, 108(11):3377–3415, 2018. 4

R. Chetty, A. Looney, and K. Kroft. Salience and taxation: Theory and evidence. *American Economic Review*, (4):1145–77, 2009. 19

R. Chetty, J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. *Quarterly Journal of Economics*, 126(4):1593–1660, 2011. 4, 20

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–2679, 2014. 4

A. Coles, L. Gray, and J. Momsen. *The Routledge Handbook of Gender and Development*. Routledge, 2015. 2

A. Cools, R. Fernández, and E. Patacchini. Girls, boys, and high achievers. Technical Report 12314, IZA Discussion Paper, 2019. 2, 23

E. Duflo, P. Dupas, and M. Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, 101(5):1739–1774, 2011. 2, 4, 23

T. L. M. Eriksen, H. S. Nielsen, and M. Simonsen. Bullying in elementary school. *The Journal of Human Resources*, 49(4):839–871, 2014. 4

M. Favara. Do dreams come true? aspirations and educational attainments of ethiopian boys and girls. *Journal of African Economies*, 26(5):561–583, 2017. 2, 4, 8, 43, 44, 48

J. Feld and U. Zölitz. Understanding peer effects: On the nature, estimation, and channels of peer effects. *Journal of Labor Economics*, (2):387–428, 2017. 20

J. C. Fruehwirth and J. Gagete-Miranda. Your peers' parents: Spillovers from parental education. *Economics of Education review*, 73:101910, 2019. 5

J. Gagete-Miranda. An aspiring friend is a friend indeed: school peers and college aspirations in brazil. Working Paper, 2020. URL http://conference.iza.org/conference_files/edu_2020/gagete_miranda_j30322.pdf. 5

D. Getik and A. Meier. Peer gender and mental health. WWZ Working Paper 2020/15, 2020. URL https://edoc.unibas.ch/78974/. 17, 20

B. Golsteyn, A. Non, and U. Zölitz. The impact of peer personality on academic achievement. *Journal of Political Economy*, 2020. In-Press. 5, 17

J. Gong, Y. Lu, and H. Song. Gender peer effects on students' academic and noncognitive outcomes: Evidence and mechanisms. *Journal of Human Resources*, 56(3):686–710, 2021. 4, 13, 23, 38

E. D. Gould, V. Lavy, and D. M. Paserman. Does immigration affect the long-term educational outcomes of natives? quasi-experimental evidence. *Economic Journal*, 119(540):1243–1269, 2009. 5

J. Guryan, K. Kroft, and M. Notowidigdo. Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, (4):34–68, 2009. 17, 60

Y. Hahn, A. Islam, E. Patacchini, and Y. Zenou. Friendship and Female Education: Evidence from a Field Experiment in Bangladeshi Primary Schools. *The Economic Journal*, 130(627):740–764, 2019. 2, 23

C. M. Hoxby. Peer effects in the classroom: Learning from gender and race variation. Working Paper, 2000. 2, 23

W. Huang, T. Li, Y. Pan, and J. Ren. Teacher Characteristics and Student Performance: Evidence from Random Teacher-Student Assignments in China. IZA Discussion Papers 14184, Institute of Labor Economics (IZA), Mar. 2021. URL https://ideas.repec.org/p/iza/izadps/dp14184.html. 19

M. Kremer and A. Holla. Improving education in the developing world: what have we learned from randomized evaluations? *Annu. Rev. Econ.*, 1(1):513–542, 2009. 8

A. B. Krueger and D. M. Whitmore. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *Economic Journal*, 111(468):1–28, 2001. 4

V. Lavy and E. Sand. On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263–279, 2018. 38

V. Lavy and A. Schlosser. Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, 3(2):1–33, 2011. 2, 3, 23, 24, 37

D. Lubotsky. Chutes or ladders? a longitudinal analysis of immigrant earnings. *Journal of Political Economy*, 115(5):820–867, 2007. 31

Ministry of Education. Curriculum framework for ethiopian education. Available at http://www.moe.gov.et/PoliciesStrategies (2020/08/28), 2009a. 7

Ministry of Education. Education statistics annual abstract. Available at http://www.moe.gov.et/EduStat (2020/08/28), 2009b. 13

P. Mouganie and Y. Wang. High-performing peers and female stem choices in school. *Journal of Labor Economics*, 38(3), 2020. 23

M. Niederle, C. Segal, and L. Vesterlund. How costly is diversity? affirmative action in light of gender differences in competitiveness. *Management Science*, 59(1):1–16, 2013. 3, 36

J. Norris. Peers, parents, and attitudes about school. *Journal of Human Capital*, 14(2):290–342, 2020. 5

K. Orkin. Are work and schooling complementary or competitive for children in rural ethiopia? a mixed-methods study. In *Childhood Poverty*, pages 298–313. Springer, 2012. 44

K. Orkin. *The effect of lengthening the school day on children's achievement in Ethiopia*. Young Lives, 2013. 7, 34

E. Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019. 25, 26, 29, 37, 62

K. Pells, P. Ogando, M. José, and P. Espinoza Revollo. Experiences of peer bullying among adolescents and associated effects on young adult outcomes: Longitudinal evidence from ethiopia, india, peru and vietnam. *Innocenti Discussion Papers*, 2016. 4

J. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73 (4):1237–1282, 2005. 24

J. Romano and M. Wolf. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, 113:38–40, 2016. 24

D. Roodman, M. Ø. Nielsen, J. MacKinnon, and M. Webb. Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal*, 19(1):4–60, 2019. 24

J. M. Rothstein. Measuring the impacts of teachers: Comment. *American Economic Review*, 107 (6):1656–1684, 2017. 4

B. Sacerdote. Experimental and quasi-experimental analysis of peer effects: Two steps forward? *Annual Review of Economics*, 6(1):253–272, 2014. doi: 10.1146/annurev-economics-071813-104217. 21

Y. Tafere and N. Chuta. Gendered trajectories of young people through school, work and marriage in ethiopia. *Young Lives Matter Working Paper*, 2016. 2

Y. Tafere and A. Pankhurst. Can children in ethiopian communities combine schooling with work? *Young Lives Matter Working Paper*, 2015. 2, 8, 38, 44

Y. Tafere and A. Tiumelissan. Slow progression: Educational trajectories of young men and women in ethiopia. *Young Lives Working Paper*, 192, 2020. 7, 8, 38

UNICEF Ethiopia. National situation analysis of children and women in ethiopia. *UNICEF Research Reports*, 2019. 2, 8, 32

W. J. Van Der Linden and R. K. Hambleton. Item response theory: Brief history, common models, and extensions. In *Handbook of Modern Item Response Theory*, pages 1–28. Springer, 1997. 10

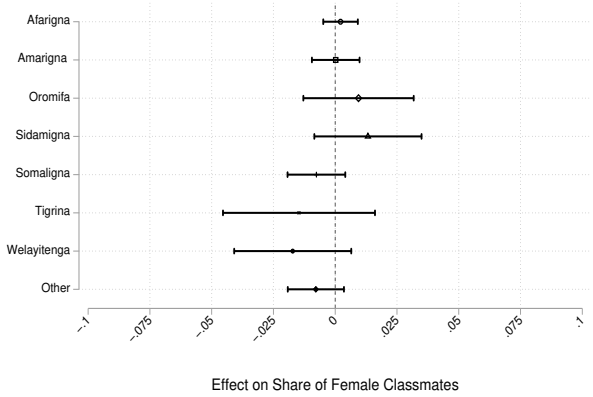
W. A. Yadete. *School management and decision-making in Ethiopian government schools*. Young Lives, 2012. 8, 39

Appendix

A Additional Tables and Figures
B Simulation for Under-Reporting in Absences

A Additional Tables and Figures

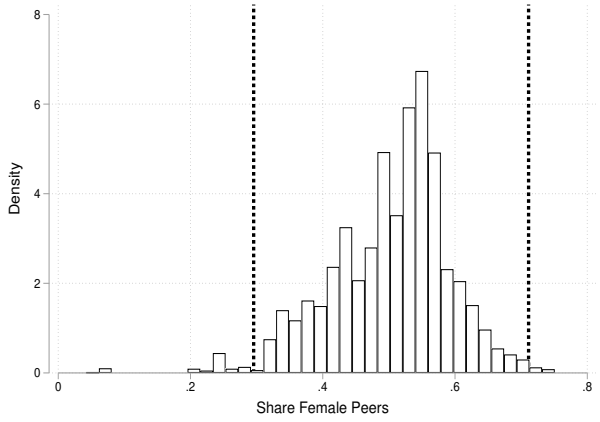
Figure A.1. Balancing Tests on Characteristics - Language and Peer Gender



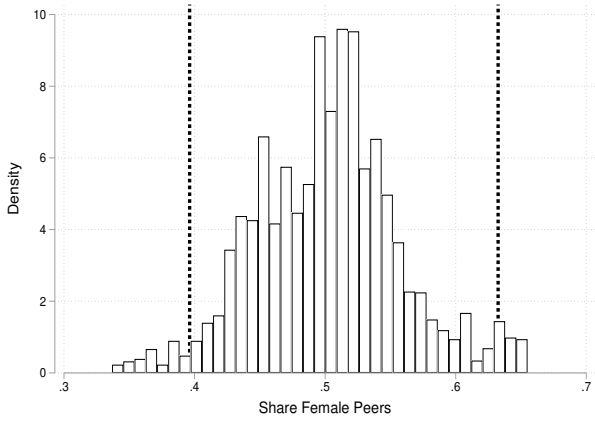
Notes: N=5077 in all cases. We regress the share of female peers on each variable (in separate regressions) on the vertical axis. The right hand side variables are dummies for specific home languages that survey respondents speak. The whiskers indicate 95% confidence intervals.

Figure A.2. Distribution of the Share of Female Classmates

(a) Raw Variation



(b) Variation Removing School Fixed Effects



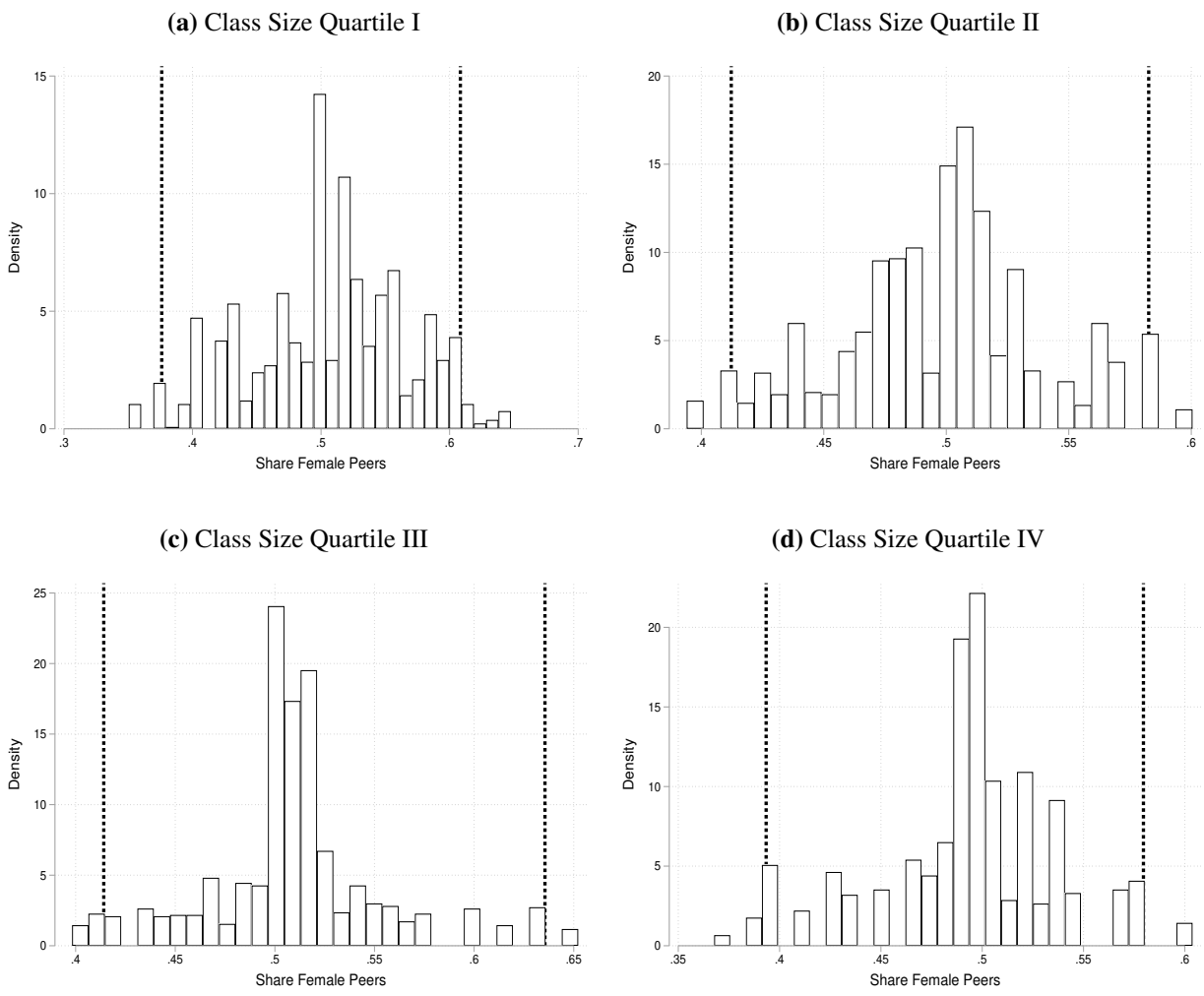
Notes: This figure presents a histogram of the share of female classmates in our selected sample. Panel (a) reports the variation in the sample, and panel (b) reports this variation after removal of school fixed effects with the sample mean added back to place it on the same scale as panel (a). Vertical lines denote the 2.5 and 97.5 percentiles.

Table A.1. Mean Differences Between Selected and Non-Selected Samples

	Selected	Non-selected	*p-value*
Outcomes			
End-Year Days Absent	5.52	6.39	0.00
End-Year Math Test Score (Std. full sample)	0.10	-0.10	0.00
End-Year Language Test Score (Std. full sample)	0.04	-0.04	0.00
Peer Variables			
Share Female Peers	0.50	0.50	0.21
Peer Start-Year Math Scores	0.09	-0.06	0.00
Peer Start-Year Language Scores	0.05	-0.04	0.00
Start-Year Test Scores			
Own Start-Year Math Scores	0.11	-0.08	0.00
Own Start-Year Language Scores	0.07	-0.05	0.00
Student Characteristics			
Female	0.51	0.50	0.15
Age (years)	11.55	11.45	0.00
Age Started School	6.68	6.97	0.00
Minority Language Spoken at Home	0.38	0.55	0.00
Number of Older Siblings	2.42	2.40	0.46
Number of Younger Siblings	1.69	1.75	0.03
Both Parents Alive	0.77	0.80	0.00
Mother Literate	0.50	0.46	0.00
Father Literate	0.57	0.60	0.00
Live with Biological Mother	0.75	0.80	0.00
Live with Father	0.58	0.64	0.00
Class Level Variables			
Start-Year Enrolled Class Size	60.20	52.40	0.00
Grade Level	4.54	4.45	0.00
Private School	0.08	0.07	0.07

Notes: Means for the selected sample and the non-selected sample are reported in columns 1 and 2. Column 3 reports the *p-value* for the statistical test of the mean differences. The outcomes end-year math and language test scores have been standardized to mean 0 and a standard deviation of 1 in the full sample prior to the analytical sample selection.

Figure A.3. Distribution of the Share of Female Classmates by Quartile of Class Size



Notes: This figure presents a histogram of the share of female classmates in our selected sample by quartile of class sizes. Vertical lines denote the 2.5 and 97.5 percentiles of the share of female class mates within the class size quartile.

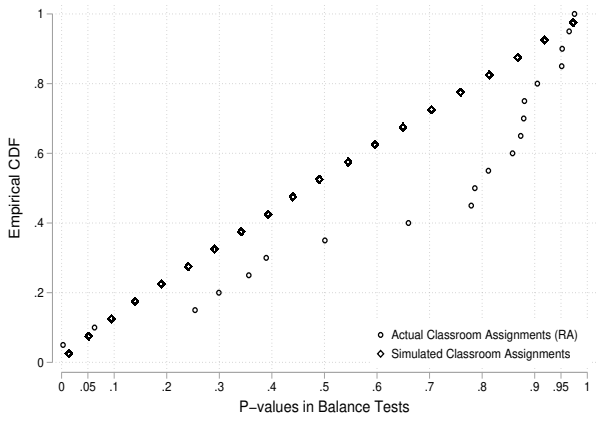
Table A.2. Share of Female Classmates and Effects on Gender

	(1)	(2)	(3)	(4)
Share Female Classmates	-0.12	-0.13	-0.14	-0.16
	(0.10)	(0.10)	(0.10)	(0.11)
School FE	Yes	Yes	Yes	Yes
School Share Female	Yes	Yes	Yes	Yes
Own-Characteristics	No	Yes	Yes	Yes
Start-Year Test Scores	No	No	Yes	Yes
Further Peer Means	No	No	No	Yes
Observations	5077	5077	5077	5077

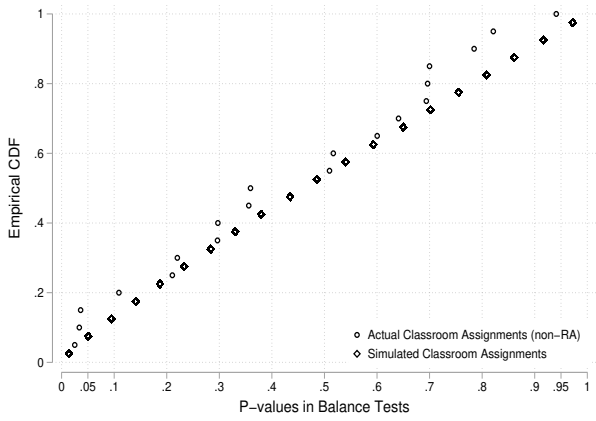
Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. All specifications are estimated on our analytical sample. The treatment variable is the share of female classmates. The dependent variable is students' own gender. In all specifications, we include the share of females at the school level to account for mechanical exclusion bias as discussed in Guryan et al. (2009) and Caeyers and Fafchamps (2020).

Figure A.4. Balance Test p-values: Simulated and Actual Class Assignments

(a) Comparisons on RA Sample



(b) Comparisons on non-RA Sample



Notes: This figure presents empirical CDF plots of the p-values from actual and pseudo-randomly class allocations within schools. The simulation tests are drawn 500 times with each of the 20 balance tests re-taken at each draw. The simulated p-value estimates are given by a bin scatter plot over 20 equally spaced bins. RA is random assignment.

Table A.3. Baseline Results - No Controls Included

	Days Absent from School		Math Test Scores		Language Test Scores	
	(1) Female	(2) Male	(3) Female	(4) Male	(5) Female	(6) Male
Share Female Classmates	-1.810**	-0.754	0.425	0.229	-0.165	0.334
	(0.797)	(0.763)	(0.408)	(0.342)	(0.316)	(0.298)
Observations	2597	2480	2597	2480	2597	2480
R^2			0.298	0.289	0.486	0.478
Controls	No	No	No	No	No	No
School FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. Columns (1) and (2) are estimated by a negative binomial regression. The treatment variable is the share of female classmates, which is the leave-one-out mean of female peers to the individual in the classroom. In Columns (1) and (2) the dependent variable is the number of days a student is absent from school over the schoolyear. In Columns (3) and (4), the dependent variable is end of the year math test scores for each student. Columns (5) and (6) show the same results for end of the year language test scores. End of year test scores are standardized. All specifications include school fixed effects but otherwise none of the baseline controls described in Table 2.

Table A.4. Baseline Outcomes and the Share of Female Peers: Mean Effects

	Days Absent	Math Scores	Language Scores
Full Sample Mean Effects			
Share Female Classmates	-1.33*	0.25	-0.06
	(0.75)	(0.24)	(0.19)

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. All specifications are estimated on our selected sample and with our baseline control set. In Column (1) the dependent variable is the number of days a student is absent from school over the schoolyear. In Column (2) the dependent variable is end of the year math test scores for each student. Column (3) show the same result for end of the year language test scores. End of year test scores are standardized.

Table A.5. Robustness to Nonlinearites in Start-Year Peer Skills

	(1) Female	(2) Male	(3) Female	(4) Male	(5) Female	(6) Male	(7) Female	(8) Male
Panel A: Days Absent								
Share Female Classmates	-1.97**	-0.62	-2.00**	-0.66	-1.81**	-0.60	-1.88**	-0.57
	(0.80)	(0.83)	(0.81)	(0.81)	(0.77)	(0.83)	(0.79)	(0.83)
Observations	2597	2480	2597	2480	2597	2480	2597	2480
Oster's δ	2.13	0.89	1.85	0.76	1.76	0.90	1.73	0.78
Panel B: Math IRT Scores								
Share Female Classmates	0.72**	-0.28	0.69**	-0.32	0.62**	-0.43	0.61**	-0.42
	(0.27)	(0.30)	(0.27)	(0.28)	(0.27)	(0.27)	(0.27)	(0.27)
Observations	2597	2480	2597	2480	2597	2480	2597	2480
R^2	0.606	0.626	0.606	0.627	0.607	0.629	0.607	0.630
Oster's δ	2.04	-0.38	1.72	-0.41	1.40	-0.52	1.35	-0.51
Panel C: Language IRT Scores								
Share Female Classmates	0.01	-0.19	0.05	-0.08	0.02	-0.13	0.02	-0.13
	(0.24)	(0.22)	(0.25)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)
Observations	2597	2480	2597	2480	2597	2480	2597	2480
R^2	0.718	0.735	0.718	0.736	0.719	0.736	0.719	0.737
Oster's δ	0.01	-0.09	0.04	-0.04	0.02	-0.06	0.02	-0.06
Peer Start-Year Test Scores	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Peer Polynomials Degree 2	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Peer Polynomials Degree 3	No	No	No	No	Yes	Yes	Yes	Yes
Peer Polynomials Degree 4	No	No	No	No	No	No	Yes	Yes

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. The treatment variable is the share of female classmates, calculated as a leave-one-out mean. In Columns (1) and (2) the dependent variable is the number of days a student is absent from school over the schoolyear. In Columns (3) and (4) the dependent variable is end of the year math test scores for each student. Columns (5) and (6) show the same result for end of the year language test scores. End of year test scores are standardized. All specifications include the base set of controls and school fixed effects as described in Table 2. Oster's delta is calculated with a $R_{\max} = 1.3 * R^2$ as suggested by Oster (2019).

Table A.6. Balance Test: Attrition

	Full Sample	Females	Males
	(1)	(2)	(3)
Peers Same Gender	-0.20***	-0.05	-0.33***
	(0.06)	(0.11)	(0.11)
Female	-0.01		
	(0.01)		
School FE	Yes	Yes	Yes
Own-Characteristics	Yes	Yes	Yes
Start-Year Test Scores	Yes	Yes	Yes
Observations	6061	3046	3007

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. Reported coefficients correspond to marginal effects obtained from probit regressions. The dependent variable is an indicator set to one if the student is observed in the start-year questionnaire, but fall out due to missing end-year test scores or days absent. Regressors are the same used in Table 2, to which we added an indicator for urban school, an indicator for having attended pre-school, and number of rooms in the house. The smaller sample size in columns (2) and (3) is due to the fact that the indicators of missing values for some variables predict the dependent variable perfectly and therefore the relevant observations and the relevant indicators are excluded from the regressions.

Table A.7. Attrition-Corrected Baseline Results - Female Sample

	Days Absent from School		Math Test Scores		Language Test Scores	
	(1) Unweighted	(2) IPW	(3) Unweighted	(4) IPW	(5) Unweighted	(6) IPW
Peer Same Gender	-1.99**	-2.31**	0.75***	0.67**	0.02	-0.23
	(0.79)	(1.03)	(0.25)	(0.28)	(0.20)	(0.26)
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Own-Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Start-Year Test Scores	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2597	2580	2597	2580	2597	2580

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. The treatment variable is the share of same-gendered classmates, calculated as a leave-one-out mean. In Columns (1) and (2) the dependent variable is the number of days a student is absent from school over the schoolyear. In Columns (3) and (4), the dependent variable is end of the year math test scores for each student. Columns (5) and (6) show the same results for end of the year language test scores. End of year test scores are standardized. Odd columns report baseline estimates for females as reported in Panel A of Table 2, even columns report attrition-corrected estimates using IPW (Inverse Probability Weighting). The weight is the inverse of the predicted probability obtained from Table A.6, column 2.

Table A.8. Attrition-Corrected Baseline Results - Male Sample

	Days Absent from School		Math Test Scores		Language Test Scores	
	(1) Unweighted	(2) IPW	(3) Unweighted	(4) IPW	(5) Unweighted	(6) IPW
Peer Same Gender	-0.64	-0.43	-0.28	-0.38	-0.12	-0.48
	(0.80)	(1.19)	(0.31)	(0.34)	(0.20)	(0.33)
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Own-Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Start-Year Test Scores	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2480	2472	2480	2472	2480	2472

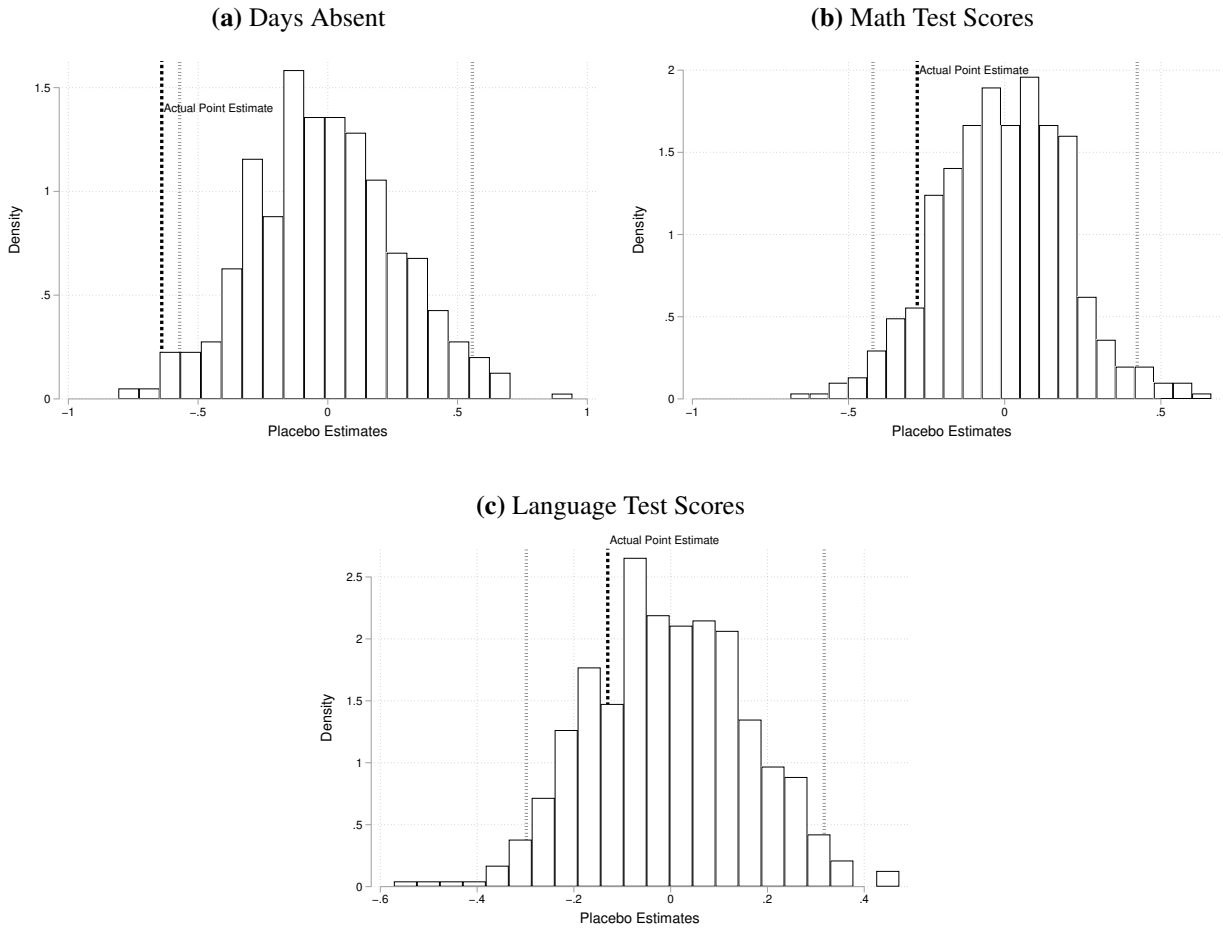
Note: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. The treatment variable is the share of same-gendered classmates, calculated as a leave-one-out mean. In Columns (1) and (2) the dependent variable is the number of days a student is absent from school over the schoolyear. In Columns (3) and (4), the dependent variable is end of the year math test scores for each student. Columns (5) and (6) show the same results for end of the year language test scores. End of year test scores are standardized. Odd columns report baseline estimates for males as reported in Panel A of Table 2, even columns report attrition-corrected estimates using IPW (Inverse Probability Weighting). The weight is the inverse of the predicted probability obtained from Table A.6, column 3.

Table A.9. Peer Gender Effects by Peer Age - Male Sample

	Days Absent			Math Scores			Language Scores		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Marginal Effects: Share Female Classmates by									
Female Peers Bottom Two Age Tertiles	-3.21			-0.06			0.23		
	(5.54)			(0.40)			(0.28)		
Female Peers Top Age Tertile	-3.26			-0.57			-0.73**		
	(6.94)			(0.41)			(0.28)		
Male Peers Bottom Two Age Tertiles		-0.95			-0.11			0.11	
		(4.36)			(0.33)			(0.20)	
Male Peers Top Age Tertile		-3.26			-0.51			0.11	
		(8.46)			(0.46)			(0.20)	
Female Peers Bottom Two Age Tertiles × Male Peers Bottom Two Tertiles			-1.40			0.09			0.39
			(4.82)			(0.41)			(0.27)
Female Peers Top Age Tertile × Male Peers Bottom Two Tertiles			6.77			-0.07			-0.42
			(13.40)			(0.35)			(0.34)
Female Peers Bottom Two Age Tertiles × Male Peers Top Tertile			-4.38			-0.52			-0.15
			(12.13)			(0.65)			(0.61)
Female Peers Top Age Tertile × Male Peers Top Tertile			-3.03			-0.60			-0.64
			(9.60)			(0.58)			(0.52)
Observations	2480	2480	2480	2480	2480	2480	2480	2480	2480

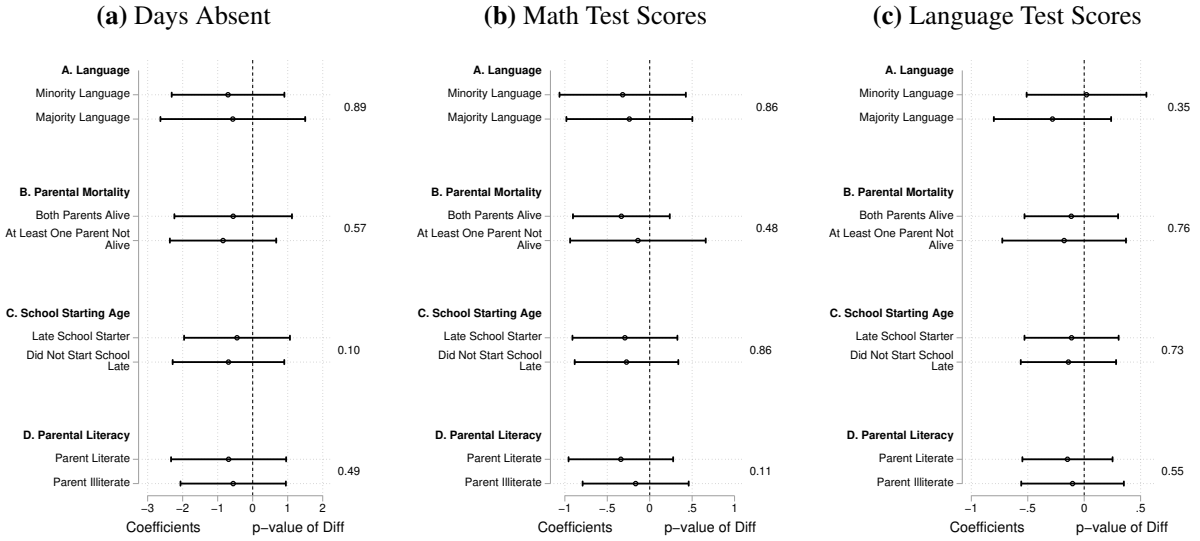
Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. We estimate our baseline specification and interact indicators for tertiles of female peers' age with indicators for tertiles of male peers' age. Estimated marginal effects correspond to the effect of the share of female classmates for each male peer age tertile/female peer age tertile combination. We restrict the sample to contain only males.

Figure A.5. Histograms of Permutation Tests: Random Re-shuffle of Students to Classrooms (Male Sample)



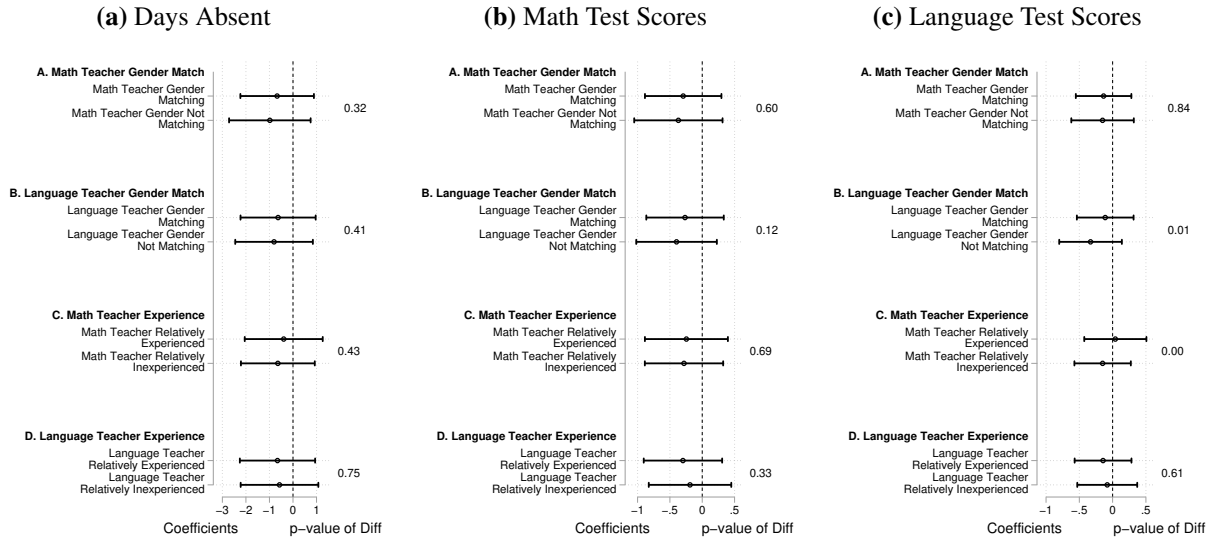
Notes: We randomly re-allocate students within schools to the classrooms, holding the number of classrooms to the number we observe in each school, and then recalculate the peer information and regression estimates. We repeat this over 500 repetitions. The true estimate is marked by the vertical dashed line and labeled, while the vertical dotted lines on the ends of the histogram represent the 2.5 and 97.5 percentile points of the simulated estimates. In panel (a), we show the histogram of the estimate from the negative binomial regression of days absent on our preferred specification from column 1 of Table 2. Panel (b) similarly reports results for math test scores and panel (c) for language test scores.

Figure A.6. Heterogeneity by Student Characteristics - Male Sample



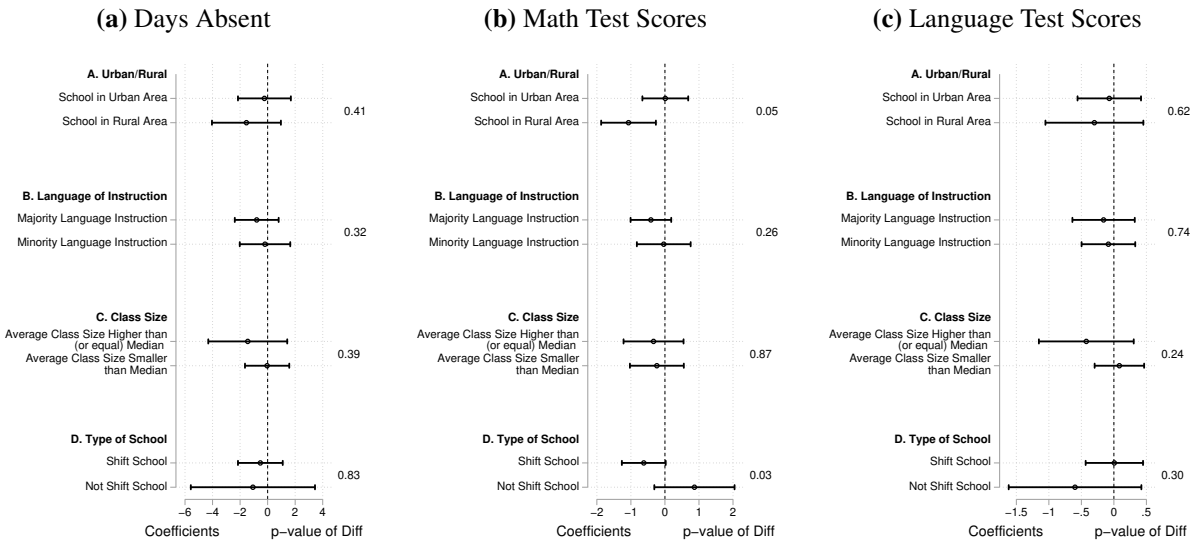
Notes:This figure presents heterogeneous effects of different subgroups on our outcomes including 95% confidence intervals clustered at the school level. We interact the share of female classmates variable with indicators of the respective student characteristics. The dependent variable is days absent in Panel (a); standardised math scores in Panel (b); and standardised language scores in Panel (c).

Figure A.7. Heterogeneity by Teacher Characteristics - Male Sample



Notes:This figure presents heterogeneous effects of different subgroups on our outcomes including 90% confidence intervals clustered at the school level. We interact the share of female classmates variable with indicators of the respective teacher characteristics. The dependent variable is days absent in Panel (a); standardised math scores in Panel (b); and standardised language scores in Panel (c).

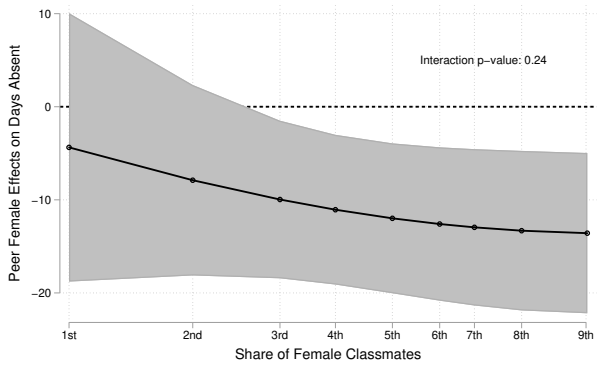
Figure A.8. Heterogeneity by School Characteristics - Male Sample



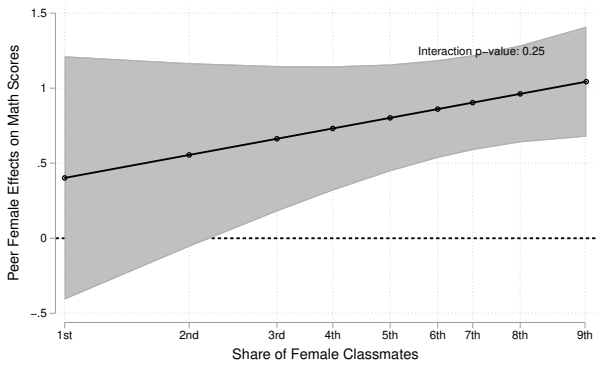
Notes:This figure presents heterogeneous effects of different subgroups on our outcomes including 95% confidence intervals clustered at the school level. We interact the share of female classmates variable with indicators of the respective school characteristics. The dependent variable is days absent in Panel (a); standardised math scores in Panel (b); and standardised language scores in Panel (c).

Figure A.9. Nonlinearity in Classmate Gender Composition: Effects on Females

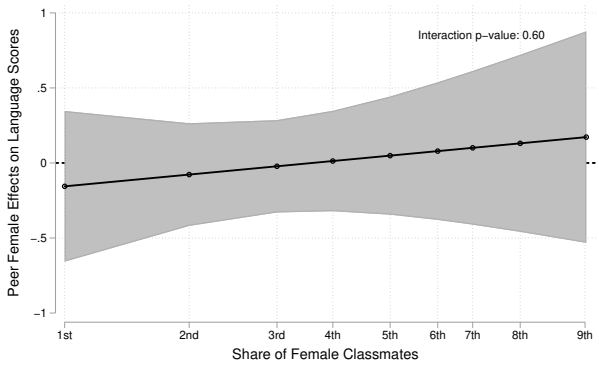
(a) Days Absent



(b) Math Test Scores



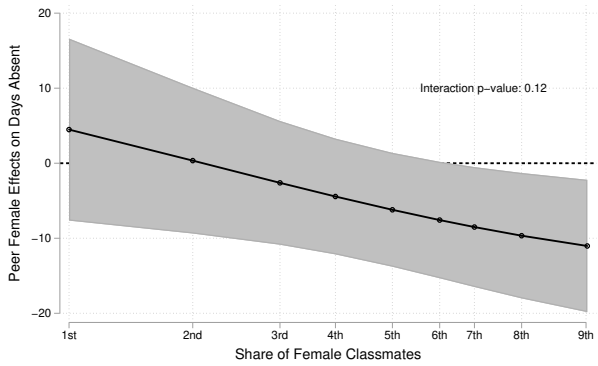
(c) Language Test Scores



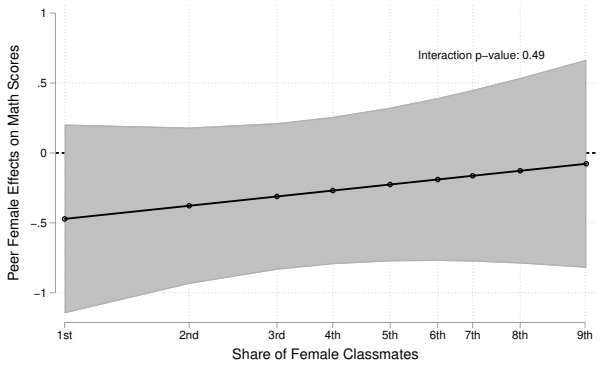
Notes: This figure presents the mean effects of the share of female classmates at deciles of peer female for females. It is based on our preferred baseline specification adding a quadratic in peer female on the subsample of females in the data. For days absent, we report the marginal effects based on the negative binomial regression. The shaded area represents 90% confidence intervals.

Figure A.10. Nonlinearity in Classmate Gender Composition: Effects on Males

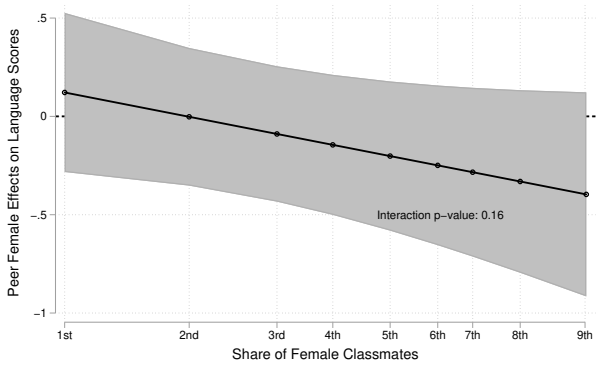
(a) Days Absent



(b) Math Test Scores



(c) Language Test Scores



Notes: This figure presents the mean effects of the share of female classmates at deciles of peer female for males. It is based on our preferred baseline specification adding a quadratic in peer female on the subsample of males in the data. For days absent, we report the marginal effects based on the negative binomial regression. The shaded area represents 95% confidence intervals.

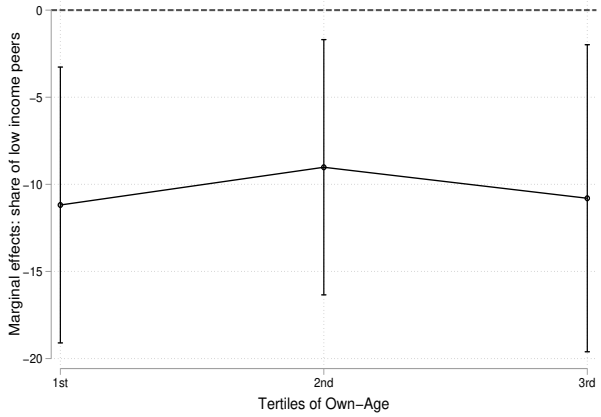
Table A.10. Summary Statistics - Teacher Motivation

	Mean	SD	Min	Max	Count
Panel A: Math Teachers					
Get through to the most difficult students	7.12	1.76	2.00	10.00	5012
Get students to learn when there is lack of support from the home	6.84	2.34	0.00	10.00	5012
Keep students on task on difficult assignments	4.99	2.78	0.00	10.00	4882
Increase students' memory of what they have been taught in previous lessons	7.91	1.75	3.00	10.00	5012
Motivate students who show low interest in schoolwork	7.94	1.25	4.00	10.00	5012
Get students to work well together	8.17	1.53	4.00	10.00	5012
Get children to do their homework	8.36	1.56	2.00	10.00	5012
Make students enjoy coming to school	7.46	2.00	0.00	10.00	5012
Get students to trust teachers	8.36	1.38	5.00	10.00	5012
Reduce school dropout	7.40	2.00	1.00	10.00	4988
Reduce school absenteeism	8.04	1.54	4.00	10.00	5012
Get students to believe they can do well in school work	8.06	1.52	4.00	10.00	5012
Panel B: Language Teachers					
Get through to the most difficult students	7.73	1.51	1.00	10.00	5003
Get students to learn when there is lack of support from the home	7.27	2.45	1.00	10.00	5003
Keep students on task on difficult assignments	4.18	3.28	0.00	10.00	5003
Increase students' memory of what they have been taught in previous lessons	8.20	1.40	0.00	10.00	5003
Motivate students who show low interest in schoolwork	8.21	1.45	2.00	10.00	5003
Get students to work well together	8.67	1.29	5.00	10.00	5003
Get children to do their homework	8.83	1.23	5.00	10.00	5003
Make students enjoy coming to school	8.12	1.44	5.00	10.00	4887
Get students to trust teachers	8.34	1.63	2.00	10.00	5003
Reduce school dropout	8.10	1.75	0.00	10.00	4979
Reduce school absenteeism	8.24	1.56	2.00	10.00	5003
Get students to believe they can do well in school work	8.34	1.49	5.00	10.00	5003

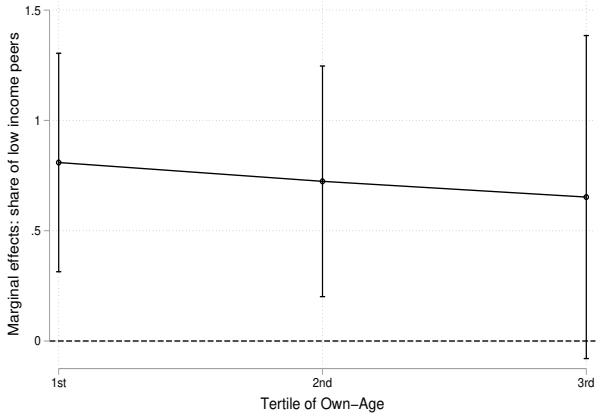
Notes: The responses indicate how much teachers agree with each statement on a scale of 0-10.

Figure A.11. Heterogeneity by Own-Age: Females

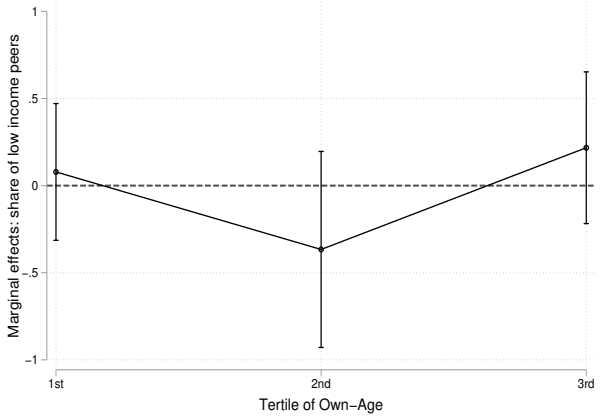
(a) Days Absent



(b) Math Test Scores



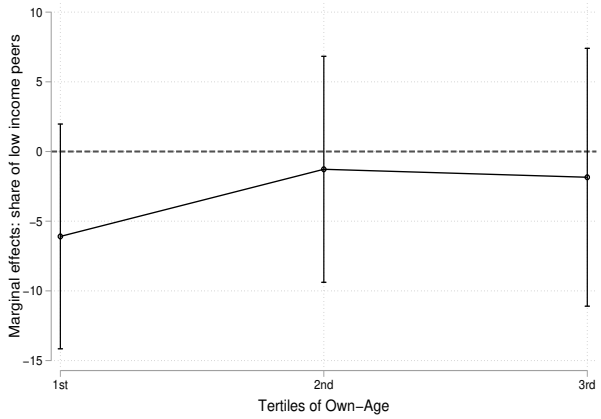
(c) Language Test Scores



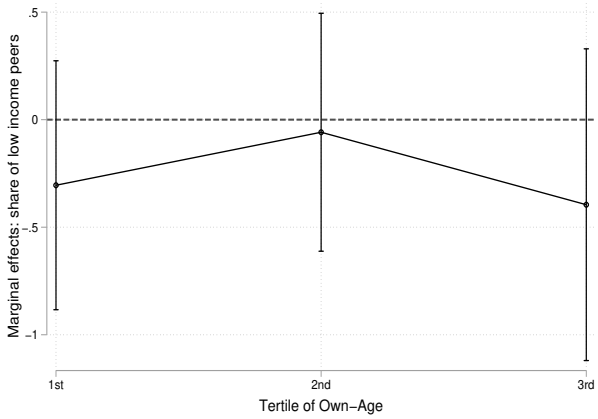
Notes: This figure presents the marginal effects for the share of female peers by terciles of students' own-age.

Figure A.12. Heterogeneity by Own-Age: Males

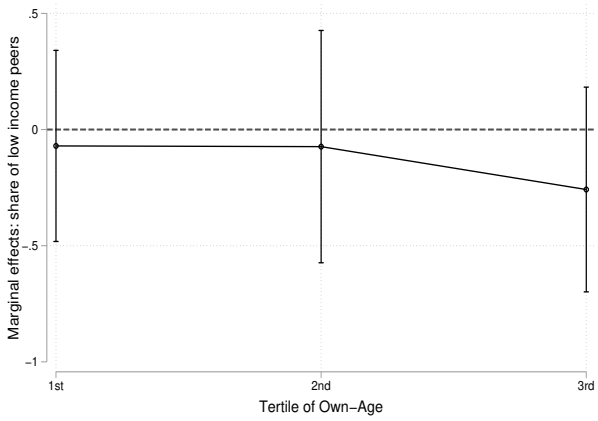
(a) Days Absent



(b) Math Test Scores



(c) Language Test Scores



Notes: This figure presents the marginal effects for the share of female peers by terciles of students' own-age.

Table A.11. Gender Differences in Previous Academic Trajectory

	Repeat Grades	Dropped Out	Pre-School
Female	-0.02	-0.02***	0.00
	(0.01)	(0.01)	(0.01)
Own-Characteristics	Yes	Yes	Yes
School and Grade FEs	Yes	Yes	Yes
Observations	5077	5077	5077

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are in parentheses and clustered at the school level. The independent variable is a dummy for female students. The dependent variable in Column (1) is an indicator for whether the student ever repeated grades. The dependent variable in Column (2) is an indicator for whether a student dropped out of school. The dependent variable in Column (3) is an indicator for whether a student attended pre-school. All specifications are estimated on our selected sample and with our baseline control set.

B Simulations for Under-Reporting in Absences

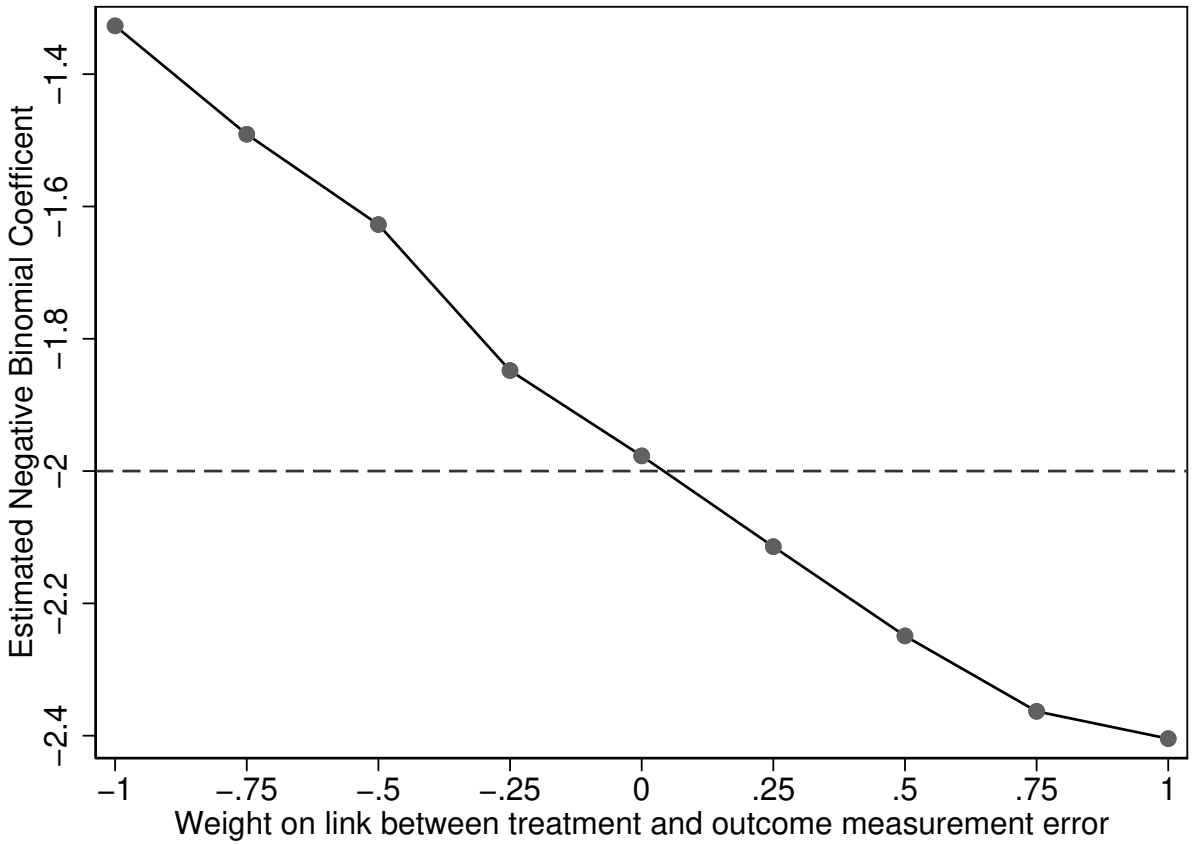
Student absences are reported at the end-year survey by teachers. One concern is that there is systematic under-reporting that could generate non-classical measurement error in our treatment effect. To help with intuition here, we simulate under-reporting in absences across 50 schools with 50 students per school. We vary the extent that the variance of this measurement error is linked with a continuous treatment that varies over schools and mimics the distribution we observe from the share of female peers, which in our data has a mean of 0.5 and an SD of 0.09. We draw our variables as follows:

$$
\begin{aligned}
y_{is} &\sim NegBin(\mu = exp(5 - 2f_s - u_i + e_i), \alpha = 1), \\
u_i &\sim |exp(1 + \gamma_k f_s) \times N(0, 1)|, \\
\gamma_k &\in \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}, \\
f_s &\sim N(0.5, 0.09), \\
e_i &\sim N(0, 1)
\end{aligned}
$$

We generate the outcome as a count variable from a negative binomial distribution and use our negative binomial coefficient estimate for the relationship between the share of low income peers and absence among females reported in Table 2 as the true effect (-2). We then vary the weight (γ_k) linking our simulated treatment to the variance of the under-reported measurement error over, collect estimates from a negative binomial regression of y_{is} on f_{is}, and then repeat 500 times. Finally, we report the mean negative binomial coefficient estimate at each γ_k in Figure B.1.

The results of our simulation show that a negative correlation between our treatment and the variance in under-reporting attenuates the effect and a positive correlation can lead to overestimation. However, the percent of bias is relatively low only rising to approximately 20% once γ_k is very high.

Figure B.1. Simulation of Measurement Error through Under-Reporting



Notes: This figure reports the mean coefficient estimate over 500 replications for each negative binomial regression at each γ_k.

Implications. One, for under-reporting to bias our estimated treatment effect requires that the variance within schools in under-reporting significantly varies as the share of female peers varies. Second, the degree of bias necessary to change our inference would require a very strong link between the variance of this measurement error and our treatment, thus we can still maintain inference in this condition, as long as the link is not too severe. Importantly, under-reporting itself will not bias our estimates without a link between its variance and the treatment.