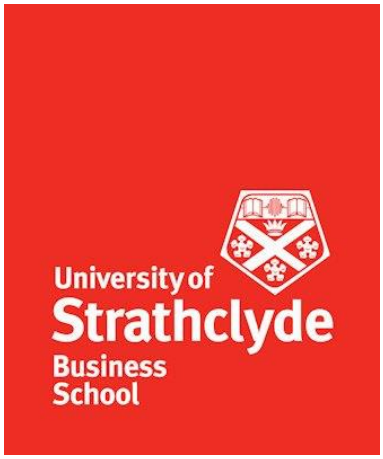
Strathclyde
Discussion Papers
in Economics



Peer Gender and Schooling: Evidence from Ethiopia

Daniel Borbely, Jonathan Norris and
Agnese Romiti

No. 21 – 4

Department of Economics
University of Strathclyde, Glasgow

Peer Gender and Schooling: Evidence from Ethiopia*

Daniel Borbely Jonathan Norris Agnese Romiti

June 2, 2021

Abstract

In this paper, we study how classmate gender composition matters for students in Ethiopia. We base our results on a unique survey of students across classrooms and schools and among those randomly assigned to class. We find a strong asymmetry: males do not and females do benefit from exposure to more female classmates with less school absence and improvement on math test scores. 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.

Keywords: Peer Effects, Gender, School Performance, Ethiopia
JEL-Codes: I21, I29, J16, J24,

*Daniel Borbely: University of Dundee, dborbely001@dundee.ac.uk; Jonathan Norris, University of Strathclyde, jonathan.norris@strath.ac.uk; Agnese Romiti, University of Strathclyde, agnese.romiti@strath.ac.uk. We thank Lukas Kiessling, Dora Gicheva, and seminar participants at the University of Strathclyde and the University of North Carolina at Greensboro for many helpful comments.

1 Introduction

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 – has received very little attention.[3] Further, 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 such as may occur 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, Ethiopia 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, Porter, and Zharkevich, 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] Within schools, conditions can vary, but class size tends to be very large, establishing an environment where teachers are spread thin and peers may form an important source of influence not

[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 Black, Devereux, and Salvanes (2013), Cools, Fernández, and Patacchini (2019), Hoxby (2000), and Lavy and Schlosser (2011).

[3] One exception is a study by Duflo, Dupas, and Kremer (2011) who 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.

[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 particularly have an expectation to maintain a good social reputation so they can secure a marriage once of age (Coles, Gray, and Momsen, 2015; Tafere and Chuta, 2016).

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 have a significant role.

We define potential mechanisms that may give rise to 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 that may 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. Where, for example, 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 degree of cost to identity utility when deviating from gender norms increases for females in the presence of males.

Recent experimental evidence confirms that beliefs and gender stereotypes can operate as a significant mechanism. When paired in groups and with a male, Bordalo et al. (2019) find females tend to significantly lower their beliefs about their ability to answer questions in gender stereotyped categories. Further, when males compete, Niederle, Segal, and Vesterlund (2013) find that females become less likely to engage in competition, and similarly, Booth and Nolen (2012) find that the presence of boys in a group has a significant role in determining girls' willingness to compete. Where

[5]Lavy and Schlosser (2011), who find similar effects across gender from exposure to a larger share of female peers in the classroom, also find a reduction in classroom disruption or violence and improved teacher-student relations consistent with this potential mechanism.

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. One, increases in the share of female peers could act to protect girls from bullying, or hostile environments, in the classroom (social interactions). In this 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); although, evidence from developed countries shows bullying has detrimental effects on educational attainment and future earnings (Brown and Taylor, 2008; Eriksen, Nielsen, and Simonsen, 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 sources for peer gender effects suggest it is important to assess the role of heterogeneity in the effects across gender.

Between symmetry and asymmetry the current empirical literature finds mixed results. Based on data in developed countries some studies have found females and males perform better in school as they are exposed to more female peers (Hoxby, 2000; Lavy and Schlosser, 2011). Others find an asymmetry where more females improve females' but not males' educational attainment (Black, Devereux, and Salvanes, 2013), that exposure in high school to high achieving males hinders females attainment but not males (Cools, Fernández, and Patacchini, 2019; Mouganie and Wang, 2020),[6] that a larger share of female classmates have similar effects across gender on math scores but only impact missed school for boys (Eren, 2017), and that more female classmates negatively impact boys' mental health (Getik and Meier, 2020).[7]

[6]There are further studies that explore the consequences of high achieving peers on men and women at university and similarly find asymmetric effects (Feld and Zölitz, 2018; Fischer, 2017). A notable exception is Anelli and Peri (2019) who find very little effects from peer gender composition in high school on choice of college major.

[7]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,

Further, mixed evidence exists among studies evaluating the effects of single-sex schooling. Among primary school students, Doris, O'Neill, and Sweetman (2013) finds that girls appear worse off relative to boys in math performance. Giardili (2019), however, finds both females and males improve in test scores, particularly where a student's gender has been disadvantaged. Among high school girls, those in single-sex classes relative to girls in coeducational classes have been documented to perform better in math and become more self-confident (Eisenkopf et al., 2015), while it appears the benefits to girls from single-sex schooling accrue to those with a strong preference for it (Jackson, 2012). More recently Jackson (2021) finds that single-sex schooling improves adolescent girls academic performance and reduces boys and girls risky behaviour through both the direct effect of peers and indirect effect of teachers.

The presence of symmetry or asymmetry for peer gender effects in the context of Ethiopia, and in general contexts where social norms and resources are different, is not clear. However, the strong gendered norms in Ethiopia suggest the potential for asymmetry where females drive any effects from classmate gender composition.

The main contributions of our study are twofold. One, we provide evidence on peer effects in an environment potentially very different from those found in the current literature. And, two, we provide new evidence on the role of classmate gender composition and mechanisms that can drive asymmetric effects across gender. We use 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]

We find that an increase in the share of female classmates decreases missed school and raises math test scores for females, while having no

students in our data are on average between 11-12 years old on the cusp of transitioning from childhood to adolescence.

[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.

effect on males. Among females, the effect sizes suggest classmate gender composition to be 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 consistent with our expectations based on the identification strategy.

We also assess a range of 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 on these dimensions, though we find suggestive evidence that the benefits from female peers are stronger as their share increases.

We further assess measures related to mechanisms that could underlie the effects we find. We focus on evidence supportive of either social interactions or shifts in teacher behaviour. First, we observe an end-of-year scale of each student's motivation and a scale of their classroom participation. Here we find females and males experience improvements on both scales from sharing the classroom with more females. 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 girls' attendance and performance. Second, we show that on a set of teacher behaviors and attitudes there is no response to the share of females in the class consistent with our effects stemming from social interactions.[9]

Third, we show that for missed school the positive effect among girls is smaller when boys in the class tend to be 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.[10] In Section 5.4.3, we discuss these results further and argue the patterns on both outcomes along classmate age are consistent with social interaction effects. While we cannot directly test for gender stereotyping and beliefs,

[9]These are teacher absences, a change in the teacher during the course of the year, and an index of teachers' self-reported belief that they can support their students learning.

[10]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.

our results are highly consistent with mechanisms via social interactions that boost motivation and allow girls to perform up to their ability.

We further 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.

Our paper contributes to the literature on peer gender composition but 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; Angrist et al., 2019; Chetty et al., 2011; Krueger and Whitmore, 2001), teacher quality (Chetty, Friedman, and Rockoff, 2014; Rothstein, 2017), effects of tracking by initial achievement conditional on teacher incentives (Duflo, Dupas, and Kremer, 2011), and peer effects over a range of dimensions. These peer dimensions include the long-run negative effect on earnings from exposure to disruptive peers (Carrell, Hoekstra, and Kuka, 2018), extensive non-linearities in peer ability effects,[11] a positive link between low-achieving Kindergarten peers and non-cognitive skills (Bietenbeck, 2020)[12] and between academic achievement and peers' persistence (Golsteyn, Non, and Zölitz, 2020), positive spillovers from friends' educational aspirations (Gagete-Miranda, 2020; Norris, 2020), and the effects of a variety of peer compositions on educational attainment.[13]

[11]See Sacerdote (2014) for a review and Feld and Zölitz (2017) for more recent work.

[12]Bietenbeck (2020) further finds this effect is driven by responses of teachers and parents that in turn boost non-cognitive skills in the classroom.

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

We add to this literature 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. Additionally, we relate to a growing literature on one particular type of peer effect – the effect of ability rank among peers on academic outcomes and behavior (Elsner and Isphording, 2017, 2018; Murphy and Weinhardt, 2020; Pagani, Comi, and Origo, forthcoming). In part, this literature suggests a source for these effects through social comparisons that in our case could potentially be stronger and more negative for girls in the presence of boys giving rise to gendered peer effects.

Alternatively, Kiessling and Norris (2020) suggest a mechanism through uncertain beliefs about ability that can be influenced through information shocks where peers in school serve as a potential source of information about ability. In our case, if information about capabilities varies with classmate gender composition and differentially across gender – for example, via gender stereotypes – then this would (i) be consistent with the belief and gender stereotyping mechanism that we outlined and (ii) further suggest asymmetric gendered peer effects.

By exploiting random assignment of students to classrooms, 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 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 popula-

tion, 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 (Favara, 2017; Tafere and Pankhurst, 2015).

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 head-teachers, and the lack of appropriate teaching facilities and infrastructure (Abebe and Woldehanna, 2013; Yadete, 2012).

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.[14] 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 prin-

[14]The survey is not representative of the population but it was designed to capture a wide range of environments within the country (Aurino, James, and Rolleston, 2014). The survey covers five out of the nine Ethiopian regions, where more than 96 percent of the population lives: Addis Ababa, Amhara, Oromia, SNNP13, and Tigray.

cipals.[15] The second survey was conducted near the end of school year.[16] 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 school. Also, included for each student are motivation and class participation scales reported by the teachers that we use of 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 pupil 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.[17] 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.[18]

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

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

[16]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, James, and Rolleston, 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.

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

[18]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.

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 "randomly/alphabetically". 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. We expect such features are unlikely to deviate within-schools, thus school fixed effects should account for this. 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 any ethno-religious differences (Ado, Gelagay, and Johannessen, 2021). In later robustness checks, we include home-language fixed effects and find no sensitivity in our results (see Section 5.2).

As we always include school fixed effects, we drop observations in schools with less than 2 classrooms.[19] Further, we drop those missing on key start-year variables, i.e. gender, the share of female classmates, and class size.[20] 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.[21]

Next, because the share of female classmates is the focus of our analysis, we use the Fisher's exact test to evaluate whether gender is balanced across classrooms within the sample. If our sample equates to a randomly assigned sample, then gender should be roughly equal across classrooms within a school. We keep observations in schools that fail to reject the null of equal distribution of gender 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.

[19]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.

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

[21]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 for our baseline set of outcomes, key variables, and controls in the selected sample. 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 7 days. Average test scores in the selected sample are higher than the mean in the full sample – at both the start and end-year surveys. We show in the appendix, Table A.1, 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. On average, classmates are evenly split between genders in both the selected and non-selected sample, while peer test scores are slightly higher than the mean in the non-selected sample.[22]

We report, in appendix Figure A.1, 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.

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, James, and Rolleston (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.

As shown in the appendix Table A.1, 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 dif-

[22]Our selected and non-selected samples have a similar proportion of private school students as well (see Table A.1), 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).

Table 1. Summary Statistics

	Mean	SD	Count
Outcomes			
End-Year Days Absent	5.52	7.32	5077
End-Year Math Test Score	-0.00	1.00	5077
End-Year Language Test Score	0.00	1.00	5077
Peer Variables			
Share Female Peers	0.50	0.09	5077
Peer Start-Year Math Scores	0.09	0.47	5027
Peer Start-Year Language Scores	0.05	0.61	5059
Start-Year Test Scores			
Own Start-Year Math Scores	0.11	0.87	5027
Own Start-Year Language Scores	0.07	0.91	5059
Student Characteristics			
Female	0.51	0.50	5077
Age (years)	11.55	1.60	5022
Age Started School	6.68	1.76	5069
Minority Language Spoken at Home	0.38	0.49	5076
Number of Older Siblings	2.42	1.86	5072
Number of Younger Siblings	1.69	1.48	5072
Both Parents Alive	0.77	0.42	5069
Mother Literate	0.50	0.50	5077
Father Literate	0.57	0.49	5077
Live with Biological Mother	0.75	0.43	5077
Live with Father	0.58	0.49	5077
Class Level Variables			
Start-Year Enrolled Class Size	60.20	15.73	5077
Grade Level	4.54	0.50	5077

Notes: The outcomes end-year math and language test scores have been standardized to mean 0 and a standard deviation of 1 in the selected sample.

ferences are small and do not represent a clear pattern of advantage or disadvantage.[23]

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.[24] The remaining controls capture characteristics about

[23]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.

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

the household – number of older and younger siblings – and about parents – having both 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.[25]

We also include two class-level controls. Class size in Ethiopia is often large (Aurino, James, and Rolleston, 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.

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:

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

[25]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.

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.[26] 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.[27] 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

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

[27]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.

bullying. Thus, we split the model by gender at the baseline, while we also report results for the full sample.[28]

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 Getik and Meier (2020), Golsteyn, Non, and Zölitz (2020), and Guryan, Kroft, and Notowidigdo (2009) 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, Kroft, and Notowidigdo (2009), to account for mechanical exclusion bias.[29] Consistent with our expectations we find no statistically significant effect.

[28]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.

[29]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.

Balance checks on additional student and teacher characteristics. In the Appendix Figure A.2 , we report point estimates and confidence intervals for each balance test on individual characteristics in panel (a) and teacher characteristics in panel (b). In each test, we regress the share of female classmates on the characteristic variable.[30] 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).

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 18 tests only 1 fails, not inconsistent with random chance. Thus, we continue to find evidence consistent with 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, Gelagay, and Johannessen, 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, 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, we

[30]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.

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.[31]

Panel (a) of Appendix Figure A.3 reports these comparisons. We report the means of the empirical CDF for the simulated p-values from 18 equally spaced bins and also the scatter plot of the empirical CDF for the actual values.[32] 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 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.

Share of female classmates 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 the share of female classmates. This follows Balestra, Eugster, and Liebert (2020), Chetty et al. (2011), and Getik and Meier (2020) and supposes that given random assignment school by class fixed effects should not represent relevant predictors of the share of female peers after accounting for school fixed effects.[33] We first obtain the residuals from regressing the share of female classmates on school fixed effects, and second, we regress these residuals on the school by class fixed effects. We also repeat this

[31]This strategy is similar to that found in Chetty, Looney, and Kroft (2009) and Huang et al. (2021).

[32]We use 18 bins because we have 18 individual balance tests in total.

[33]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.

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.

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.[34] 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 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.[35]

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%

[34] We use classmates and peers interchangeably.

[35] In robustness checks, we consider a wide range of additional controls.

of the share of female peers.[36] 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.[37]

To put our findings into the context of similar studies based on richer countries, we consider Lavy and Schlosser (2011), the closest paper examining, among other outcomes, the effect of female peers on math test scores among 5^{th} graders in Israel.[38] The magnitude of our effects are bigger than in Lavy and Schlosser (2011), though they consider peers at the level of school grade as opposed to classroom level peers as in our case.[39] Our larger effects suggest that the role of peers 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,[40] 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 peers in a school environment that is richer than the one we study.

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, rejecting the null of equality. These results strongly suggest that in our sample males do not

[36]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.

[37]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.

[38]Duflo, Dupas, and Kremer (2011) is the only paper analysing peer effects in a developing country context, however their paper focuses on peer quality and does not examine peer gender.

[39]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.

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

and females do benefit from exposure to more female classmates. This is inline 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.

In the Appendix Table A.3, 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.[41] 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, Gelbach, and Miller, 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).[42] We only recored small increases in the *p-values* and maintain a 5% significance level for both days absent and math test scores.

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.[43] 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,

[41] 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.

[42] 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/`.

[43] 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 $R_{\max} = 1.3 * R^2$.

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. Share of female classmates is the leave-one-out mean of female peers to the individual in the classroom. 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.

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 B.1 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. We again expand the controls with potentially relevant dimensions. In the Appendix Table B.2, 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 A.2. 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. This set, not including school fixed effects, contains 115 controls. We then use a post-double selection (PDS) lasso (Belloni, Chernozhukov, and Hansen, 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.[44] Inference is not valid on the selected controls,

[44]We do not penalize school fixed effects because accounting for shocks at the level of random assignment, the school, is still important.

however, Belloni, Chernozhukov, and Hansen (2014) show that it remains valid for the treatment, in our case the share of female peers.

We find evidence highly consistent with our baseline results.[45] As we iterate through specifications, the effects generally remain similar in size, and always qualitatively consistent, across outcomes and gender. In particular, 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 the Appendix Figure B.1, 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 and math test scores), they fall entirely outside the distribution of simulated effects.[46]

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 students to classrooms.

[45] 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.

[46] We report our results on males in Figure B.2. 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.

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 the Appendix Figure C.1, for females, and Figure C.2 for males. Focusing on females, in general the effects tend to be similar across categories.[47] 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. Nevertheless, we do not find these heterogeneities on math test scores and caution against making strong conclusions based on only this result.

Teacher characteristics. We report similar heterogeneity results by a set of teacher characteristics for females and males in the Appendix Figures C.3 and C.4. We distinguish between math and language teachers' characteristics, but control for the possibility that the same teacher might

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

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 Appendix Figures C.5 and C.6, we report heterogeneity by a set of school characteristics, for females and males, respectively. 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 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 C.7 and C.8, 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 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, Segal, and Vesterlund, 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 3, we report estimated effects from the share of female classmates on end-of-year motivation and participation.[48] For each, we report the estimated effect among all students and then split by gender.

Table 3. 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. 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.[49] The point estimates are larger for females but not statistically different from those for males.[50]

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 in Israeli data, or less competitive social interactions if females tend to compete less stringently. However, these positive benefits

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

[49]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%.

[50]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.

for boys do not translate into improvements on missed schooling and test scores.

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.[51] Thus, we now turn to investigate some observable teacher behaviors and attitudes, which 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 4, 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. 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.

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

[51] 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).

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 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.4 in the Appendix. 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.

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. For motivation, since we observe this at the start-year survey we cannot rule out that exposure to more females over the school year shifts teacher beliefs and attitudes. Nevertheless, on the end-year measures for teacher absences and whether the teacher changed,

[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.

Table 4. 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 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.

we find no effects. In general, teachers in Ethiopia face many other factors and constraints that likely drive their behavior. For example, teacher incentives and compensation might be poor, and teaching facilities are often inadequate (Abebe and Woldehanna, 2013; Yadete, 2012). Thus, our results here, are consistent with peer gender effects acting directly through social interaction mechanisms in a setting of large classrooms and teachers equipped with poor resources and incentives.

5.4.3 Heterogeneity by Classmate Age

Next, we address whether the age of classmates varies the peer gender effect. We suggested two feasible mechanisms for peer gender effects within social interactions: (i) shifts in girls' beliefs about capabilities and (ii) shifts in protection from bullying. 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 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.[53]

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.[54] We then add to our baseline specification a full set of interactions between these indicators and the share of female classmates.[55] 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 5, 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 D.1.

Our results for girls, in Table 5, suggest some important heterogeneities. For days absent, girls benefit, significantly reducing their absences, from

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

[54]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.

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

exposure to more girls regardless of whether 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. We find this consistent with a social interaction mechanism. An example would be 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.

Table 5. Peer Gender Effects by Peer Age - Female Sample

	Days Absent		Math Scores		Language Scores	
Female Peers Age Tertile	Bottom Two	Oldest	Bottom Two	Oldest	Bottom Two	Oldest
Male Peers = Bottom Two	-13.57**	-13.93**	1.12***	-0.40	0.24	-0.14
	(5.53)	(6.44)	(0.32)	(0.47)	(0.24)	(0.31)
Male Peers = Oldest	-5.25	-7.37	1.30***	0.13	0.15	-0.36
	(10.21)	(8.48)	(0.24)	(0.40)	(0.38)	(0.41)
Observations	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. 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 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. Thus, we believe these point toward mechanisms driven by social interaction.

In general, the peer gender effect among females exhibits strong patterns of heterogeneity consistent with social interaction mechanisms. Turning to boys, reported in Appendix Table D.1, 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.

5.4.4 The Effect on Bullying

As discussed above, an increase in the share of female classmates could act as a protection mechanism to alleviate bullying in the classroom. Unfortunately, the school survey we use for our analysis does not contain information on bullying behaviour. However, the Younger Cohort (YC) of the Young Lives longitudinal survey does have a variable indicating whether a student was ever bullied by their peers. This survey tracks a smaller group of schoolchildren over time and a small sub-set of the children included in the YC study are also included in the school survey.

We match the data on bullying from Wave 4 of the longitudinal YC survey with our school survey.[56] We then test whether the share of female classmates has any impact on the probability of being bullied. We estimate this (i) only controlling for gender, (ii) using the full set of baseline controls, and (iii) using a post-double selection lasso model to select the controls that are the best predictors of both the outcome and classmate gender composition. Results are reported in the Appendix Table D.2.

Overall, the results indicate that the share of female classmates is negatively associated with bullying for females, but positively associated with bullying for males. Although none of these associations are significant, they provide suggestive evidence that girls face less bullying when they have a higher share of female classmates. Note, that using the YC data for this analysis has led to a substantial reduction in sample size and we simply may not have the power to detect significant effects.[57]

Our findings here, however, go in the expected direction if the protection mechanism indeed plays a part in helping girls achieve better outcomes at school.

Putting together our results on mechanisms, our evidence suggests that social interactions drive the effects from the share of female peers. Further, they point to important sources on motivation and participation in class

[56]Wave 4 of the YC survey took place in the academic year 2013/14, which corresponds to the next academic year of the school survey.

[57]Although our bullying measure is not specific in terms of the timing of the event, as it accounts for whether someone has ever been bullied, the fact that we can't be sure that the treatment preceded the outcome is not problematic in our setting where we have randomization of peers. Any event preceding the measure of peers should be randomly distributed across students.

and along the age of classmates, while indicating the absence of effects on observable teacher behaviors.

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 report the marginal effects corresponding to these categories in Table 6. Our results are disaggregated across gender and across different types of child work.[58]

[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.

Table 6. 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
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 and c) an indicator for whether a student is involved in more hours of domestic work than the median. 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.

It is clear from Table 6 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

[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.

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.

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 our observable 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.

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 (Alan, Ertac, and Mumcu, 2018; Favara, 2017).

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

Abebe, Workneh and Tassew Woldehanna (2013). *Teacher training and development in Ethiopia: Improving education quality by developing teacher skills, attitudes and work conditions*. Young Lives.

Ado, Derib, Almaz Wasse Gelagay, and Janne Bondi Johannessen (2021). "The languages of Ethiopia". In: *Grammatical and Sociolinguistic Aspects of Ethiopian Languages* 48, p. 1.

Akerlof, George and Rachel Kranton (2000). "Economics and Identity". In: *The Quarterly Journal of Economics* 115.3, pp. 715–753.

Alan, Sule, Seda Ertac, and Ipek Mumcu (2018). "Gender stereotypes in the classroom and effects on achievement". In: *Review of Economics and Statistics* 100.5, pp. 876–890.

Anelli, Massimo and Giovanni Peri (2019). "The Effects of High School Peers' Gender on College Major, College Performance and Income". In: *Economic Journal* 129.618, pp. 553–602.

Angrist, Joshua (2014). "The Perils of Peer Effects". In: *Labour Economics*, pp. 98–108.

Angrist, Joshua D. and Victor Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". In: *Quarterly Journal of Economics* 114.2, pp. 533–575.

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany (2019). "Maimonides' Rule Redux". In: *American Economic Review: Insights* 1.3, pp. 309–324.

Aurino, Elisabetta, Zoe James, and Caine Rolleston (2014). *Young Lives Ethiopia School Survey 2012-13: Data Overview Report*. Tech. rep. Working Paper 134. Young Lives.

Balestra, Simone, Beatrix Eugster, and Helge Liebert (2020). "Peers with Special Needs". In: *Review of Economics and Statistics*.

Bau, Natalie, Martin Rotemberg, Manisha Shah, and Bryce Steinberg (2020). *Human Capital Investment in the Presence of Child Labor*. Tech. rep. National Bureau of Economic Research.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen (2012). "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain". In: *Econometrica* 80.6, pp. 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on Treatment Effects after Selection Among High-Dimensional Controls". In: *The Review of Economic Studies* 81.2, pp. 608–650.

Bertrand, Marianne and Jessica Pan (2013). "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior". In: *American Economic Journal: Applied Economics* 5.1, pp. 32–64.

Bietenbeck, Jan (2020). "The Long-Term Impacts of Low-Achieving Childhood Peers: Evidence from Project STAR". In: *Journal of the European Economic Association* 18.1, pp. 392–426.

Bifulco, Robert, Jason M. Fletcher, Sun Jung Oh, and Stephen L. Ross (2014). "Do high school peers have persistent effects on college attainment and other life outcomes?" In: *Labour Economics* 29, pp. 83–90.

Bifulco, Robert, Jason M. Fletcher, and Stephen L. Ross (2011). "The Effect of Classmate Characteristics on Post-secondary Outcomes: Evidence from the Add Health". In: *American Economic Journal: Economic Policy* 3.1, pp. 25–53.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes (2013). "Under Pressure? The Effect of Peers on Outcomes of Young Adults". In: *Journal of Labor Economics* 31.1, pp. 119–153.

Booth, Alison and Patrick Nolen (2012). "Choosing to compete: How different are girls and boys?" In: *Journal of Economic Behavior & Organization* 81.2, pp. 542–555.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer (2019). "Beliefs about Gender". In: *American Economic Review* 109.3, pp. 739–73.

Boyden, Jo, Catherine Porter, and Ina Zharkevich (2020). "Balancing school and work with new opportunities: changes in children's gendered time use in Ethiopia (2006–2013)". In: *Children's Geographies*, pp. 1–14.

Brown, Sarah and Karl Taylor (2008). "Bullying, education and earnings: Evidence from the National Child Development Study". In: *Economics of Education Review* 27.4, pp. 387–401.

Caeyers, Bet and Marcel Fafchamps (2020). *Exclusion Bias in the Estimation of Peer Effects*. Tech. rep. DP14386. CEPR Discussion Paper Series.

Cameron, Colin, Jonah Gelbach, and Douglas Miller (2008). "Bootstrap-based improvements for inference with clustered errors". In: *The Review of Economics and Statistics* 90.3, pp. 414–427.

Cameron, Colin and Douglas Miller (2015). “A Practitioner’s Guide to Cluster-Robust Inference”. In: *Journal of Human Resources* 50.2, pp. 317–372.

Carrell, Scott E., Mark Hoekstra, and Elira Kuka (2018). “The Long-Run Effects of Disruptive Peers”. In: *American Economic Review* 108.11, pp. 3377–3415.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011). “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star”. In: *Quarterly Journal of Economics* 126.4, pp. 1593–1660.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014). “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood”. In: *American Economic Review* 104.9, pp. 2633–2679.

Chetty, Raj, Adam Looney, and Kory Kroft (2009). “Salience and Taxation: Theory and Evidence”. In: *American Economic Review* 4, pp. 1145–77.

Coles, Anne, Leslie Gray, and Janet Momsen (2015). *The Routledge Handbook of Gender and Development*. Routledge.

Cools, Angela, Raquel Fernández, and Eleonora Patacchini (2019). *Girls, Boys, and High Achievers*. Tech. rep. 12314. IZA Discussion Paper.

Doris, Aedín, Donal O’Neill, and Olive Sweetman (2013). “Gender, single-sex schooling and maths achievement”. In: *Economics of Education Review* 35, pp. 104–119.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011). “Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya”. In: *American Economic Review* 101.5, pp. 1739–1774.

Eisenkopf, Gerald, Zohal Hessami, Urs Fischbacher, and Heinrich Ursprung (2015). “Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland”. In: *Journal of Economic Behavior & Organization* 115, pp. 123–143.

Elsner, Benjamin and Ingo E. Isphording (2017). “A Big Fish in a Small Pond: Ability Rank and Human Capital Investment”. In: *Journal of Labor Economics* 35.3, pp. 787–828.

— (2018). “Rank, Sex, Drugs, and Crime”. In: *Journal of Human Resources* 52.2, pp. 356–381.

Eren, Ozkan (2017). “Differential Peer Effects, Student Achievement, and Student Absenteeism: Evidence From a Large-Scale Randomized Experiment”. In: *Demography* 54.2, pp. 745–773.

Eriksen, Tine Louise Mundbjerg, Helena Skyt Nielsen, and Marianne Simonsen (2014). “Bullying in Elementary School”. In: *The Journal of Human Resources* 49.4, pp. 839–871.

Favara, Marta (2017). “Do dreams come true? Aspirations and educational attainments of Ethiopian boys and girls”. In: *Journal of African Economies* 26.5, pp. 561–583.

Feld, Jan and Ulf Zölitz (2017). “Understanding Peer Effects: On the Nature, Estimation, and Channels of Peer Effects”. In: *Journal of Labor Economics* 2, pp. 387–428.

— (2018). “Peers from Venus and Mars – higher-achieving men foster gender gaps in major choice and labor market outcomes”. Working Paper.

Fischer, Stefanie (2017). “The downside of good peers: How classroom composition differentially affects men’s and women’s STEM persistence”. In: *Labour Economics* 46, pp. 211–226.

Fruehwirth, Jane Cooley and Jessica Gagete-Miranda (2019). “Your peers’ parents: Spillovers from parental education”. In: *Economics of Education review* 73, p. 101910.

Gagete-Miranda, Jessica (2020). “An aspiring friend is a friend indeed: school peers and college aspirations in Brazil”. Working Paper.

Getik, Demid and Armando Meier (2020). “Peer Gender and Mental Health”. WWZ Working Paper 2020/15.

Giardili, Soledad (2019). *Single-Sex Primary Schools and Student Achievement: Evidence from Admission Lotteries*. WorkingPaper.

Golsteyn, Bart, Arjan Non, and Ulf Zölitz (2020). “The Impact of Peer Personality on Academic Achievement”. In: *Journal of Political Economy*. In-Press.

Gould, Eric D., Victor Lavy, and Daniele M. Paserman (2009). “Does Immigration Affect the Long-Term Educational Outcomes of Natives? Quasi-Experimental Evidence”. In: *Economic Journal* 119.540, pp. 1243–1269.

Guryan, Jonathan, Kory Kroft, and Mathew Notowidigdo (2009). “Peer Effects in the Workplace: Evidence from Random Groupings in Pro-

fessional Golf Tournaments". In: *American Economic Journal: Applied Economics* 4, pp. 34–68.

Hoxby, Caroline M. (2000). "Peer Effects in the Classroom: Learning from Gender and Race Variation". Working Paper.

Huang, Wei, Teng Li, Yinghao Pan, and Jinyang Ren (Mar. 2021). *Teacher Characteristics and Student Performance: Evidence from Random Teacher-Student Assignments in China*. IZA Discussion Papers 14184. Institute of Labor Economics (IZA).

Jackson, C Kirabo (2021). "Can Introducing Single-Sex Education into Low-Performing Schools Improve Academics, Arrests, and Teen Motherhood?" In: *Journal of Human Resources* 56.1, pp. 1–39.

Jackson, Kirabo (2012). "Single-sex schools, student achievement, and course selection: Evidence from rule-based student assignments in Trinidad and Tobago". In: *Journal of Public Economics* 96.1-2, pp. 173–187.

Kiessling, Lukas and Jonathan Norris (2020). "The Long-Run Effects of Peers on Mental Health". Working Paper.

Kremer, Michael and Alaka Holla (2009). "Improving education in the developing world: what have we learned from randomized evaluations?" In: *Annu. Rev. Econ.* 1.1, pp. 513–542.

Krueger, Alan B. and Diane M. Whitmore (2001). "The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star". In: *Economic Journal* 111.468, pp. 1–28.

Lavy, Victor and Edith Sand (2018). "On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases". In: *Journal of Public Economics* 167, pp. 263–279.

Lavy, Victor and Analia Schlosser (2011). "Mechanisms and Impacts of Gender Peer Effects at School". In: *American Economic Journal: Applied Economics* 3.2, pp. 1–33.

Ministry of Education (2009a). *Curriculum Framework for Ethiopian Education*. Available at `http://www.moe.gov.et/PoliciesStrategies` (2020/08/28).

— (2009b). *Education Statistics Annual Abstract*. Available at `http://www.moe.gov.et/EduStat` (2020/08/28).

Mouganie, Pierre and Yaojing Wang (2020). "High-Performing Peers and Female STEM Choices in School". In: *Journal of Labor Economics* 38.3.

Murphy, Richard and Felix Weinhardt (2020). “Top of the Class: The Importance of Ordinal Rank”. In: *Review of Economic Studies* 87.6, pp. 2777–2826.

Niederle, Muriel, Carmit Segal, and Lise Vesterlund (2013). “How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness”. In: *Management Science* 59.1, pp. 1–16.

Norris, Jonathan (2020). “Peers, Parents, and Attitudes about School”. In: *Journal of Human Capital* 14.2, pp. 290–342.

Orkin, Kate (2012). “Are work and schooling complementary or competitive for children in rural Ethiopia? A mixed-methods study”. In: *Childhood Poverty*. Springer, pp. 298–313.

— (2013). *The effect of lengthening the school day on children’s achievement in Ethiopia*. Young Lives.

Oster, Emily (2019). “Unobservable Selection and Coefficient Stability: Theory and Evidence”. In: *Journal of Business & Economic Statistics* 37.2, pp. 187–204.

Pagani, Laura, Simona Comi, and Federica Origo (forthcoming). “The Effect of School Rank on Personality Traits”. In: *Journal of Human Resources*.

Pells, Kirrily, Portela Ogando, Maria José, and Patricia Espinoza Revollo (2016). “Experiences of Peer Bullying among Adolescents and Associated Effects on Young Adult Outcomes: Longitudinal Evidence from Ethiopia, India, Peru and Vietnam”. In: *Innocenti Discussion Papers*.

Romano, Joseph and Michael Wolf (2005). “Stepwise Multiple Testing as Formalized Data Snooping”. In: *Econometrica* 73.4, pp. 1237–1282.

— (2016). “Efficient Computation of Adjusted p-values for Resampling-Based Stepdown Multiple Testing”. In: *Statistics & Probability Letters* 113, pp. 38–40.

Roodman, David, Morten Ørregaard Nielsen, James MacKinnon, and Matthew Webb (2019). “Fast and wild: Bootstrap inference in Stata using boottest”. In: *The Stata Journal* 19.1, pp. 4–60.

Rothstein, Jesse M. (2017). “Measuring the Impacts of Teachers: Comment”. In: *American Economic Review* 107.6, pp. 1656–1684.

Sacerdote, Bruce (2014). “Experimental and Quasi-Experimental Analysis of Peer Effects: Two Steps Forward?” In: *Annual Review of Economics* 6.1, pp. 253–272.

Tafere, Yisak and Nardos Chuta (2016). “Gendered trajectories of young people through school, work and marriage in Ethiopia”. In: *Young Lives Matter Working Paper*.

Tafere, Yisak and Alula Pankhurst (2015). “Can children in Ethiopian communities combine schooling with work?” In: *Young Lives Matter Working Paper*.

Tafere, Yisak and Agazi Tiumelissan (2020). “Slow Progression: Educational Trajectories of Young Men and Women in Ethiopia”. In: *Young Lives Working Paper* 192.

UNICEF Ethiopia (2019). “National Situation Analysis of Children and Women in Ethiopia”. In: *UNICEF Research Reports*.

Van Der Linden, Wim J and Ronald K Hambleton (1997). “Item Response Theory: Brief History, Common Models, and Extensions”. In: *Handbook of Modern Item Response Theory*. Springer, pp. 1–28.

Yadete, Workneh Abebe (2012). *School management and decision-making in Ethiopian government schools*. Young Lives.

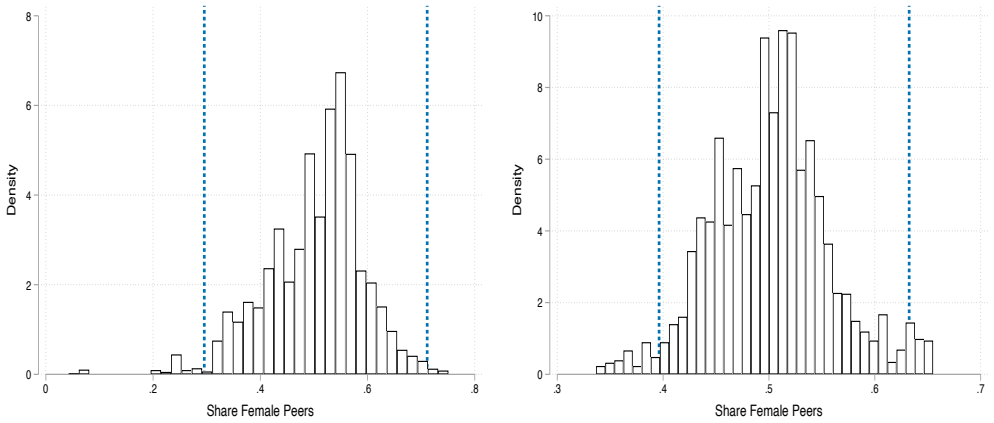
Appendix – For Online Publication

A Additional Tables and Figures
B Tables and Figures for Robustness Checks
C Heterogeneity Results
D Additional Results for Mechanisms

A Additional Tables and Figures

Figure A.1. 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.

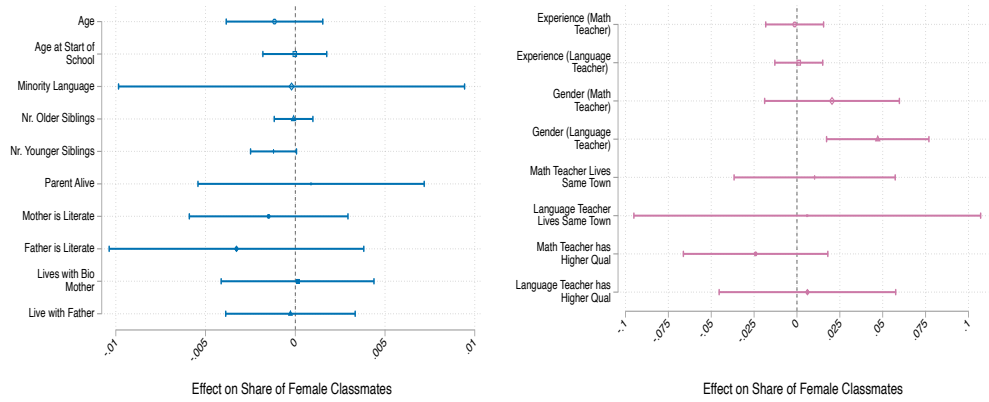
Table A.2. Share of Female Classmates and Effects on Gender

	(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 specification are estimated on our analytical sample. In all specifications, we include the share of females at the school level to account for mechanical exclusion bias as discussed in Guryan, Kroft, and Notowidigdo (2009) and Caeyers and Fafchamps (2020).

Figure A.2. Balancing Tests on Characteristics

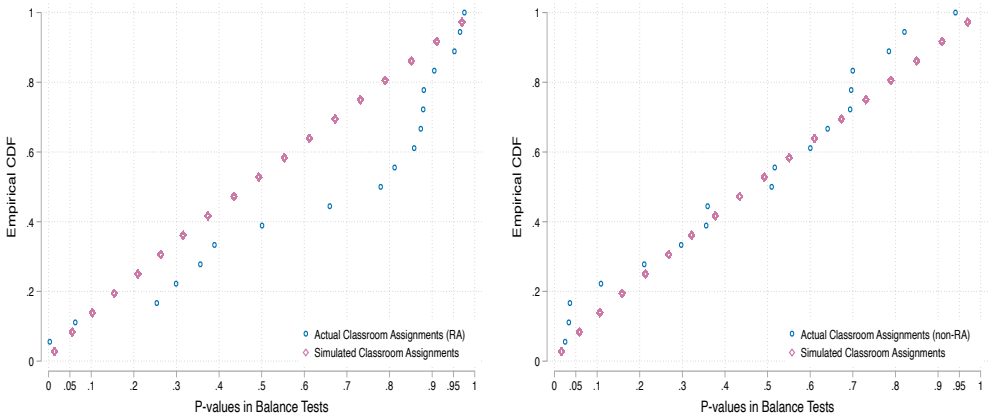
(a) Student Characteristics **(b)** Teacher Characteristics



Notes: N=5077 in all cases. We regress the share of female peers on each variable on the vertical axis. In panel (a) the right hand side variables are student characteristics. In panel (b) the right hand side variables are teacher characteristics. The whiskers indicate 95% confidence intervals.

Figure A.3. 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 18 balance tests re-taken at each draw. The simulated p-value estimates are given by a bin scatter plot over 18 equally spaced bins. RA is random assignment.

Table A.3. 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 specification are estimated on our selected sample and with our baseline control set.

Table A.4. Summary Statistics - Teacher Motivation

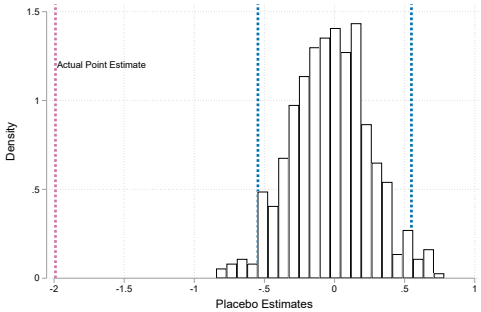
	Mean	SD	Min	Max	Count
Panel A: Math Teachers					
Get through to the most difficult students	7.12	1.76	2.00	10.00	5012
Get students to learn when there is lack of support from the home	6.84	2.34	0.00	10.00	5012
Keep students on task on difficult assignments	4.99	2.78	0.00	10.00	4882
Increase students' 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.

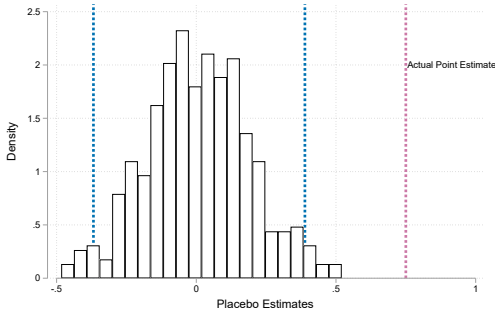
B Tables and Figures for Robustness Checks

Figure B.1. Histograms of Permutation Tests: Random Re-shuffle of Students to Classrooms (Female Sample)

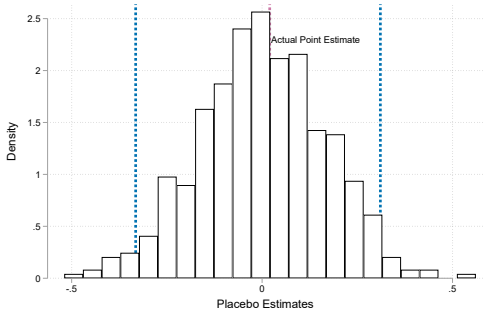
(a) Days Absent



(b) Math Test Scores



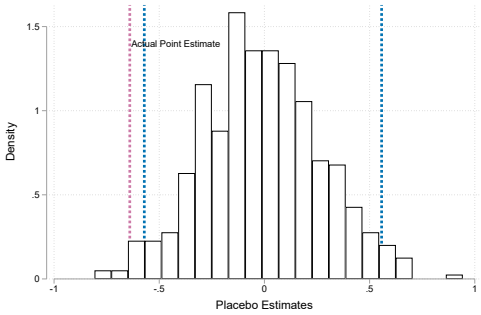
(c) Language Test Scores



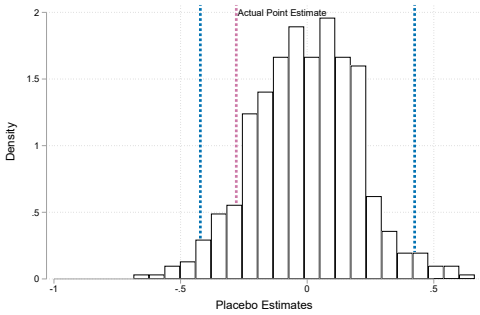
Notes: We randomly re-allocate students within schools to the classrooms, holding the number of classrooms to the number we observe in each school, and then recalculate the peer information and regression estimates. We repeat this over 500 repetitions. The true estimate is marked by the vertical dashed line and labeled, while the vertical dashed 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 B.2. Histograms of Permutation Tests: Random Re-shuffle of Students to Classrooms (Male Sample)

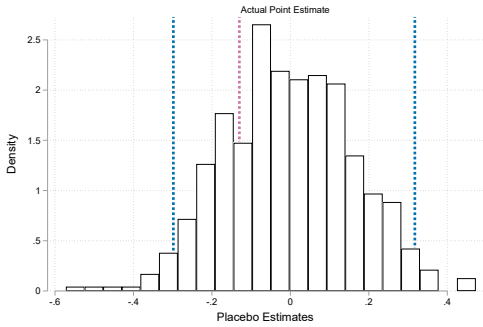
(a) Days Absent



(b) Math Test Scores



(c) Language Test Scores



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 dashed 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.

Table B.1. 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. 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 B.2. Additional Specifications and High Dimensional Controls

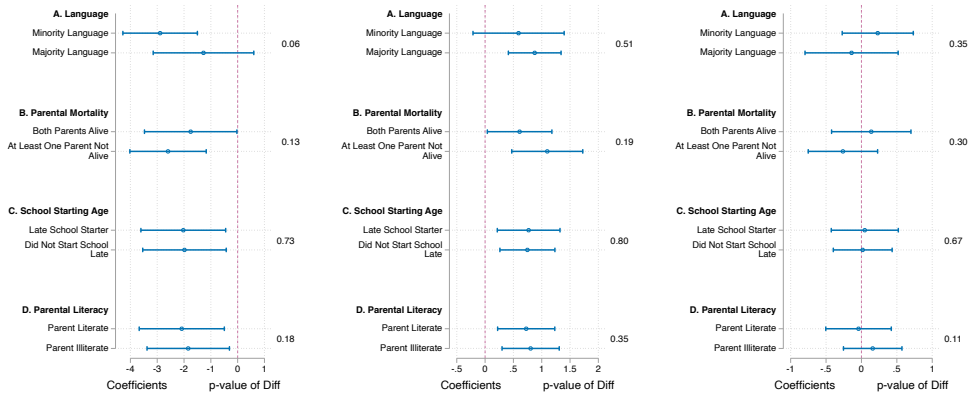
	OLS (unpenalized)						PDS Lasso	
	(1) Female	(2) Male	(3) Female	(4) Male	(5) Female	(6) Male	(7) Female	(8) Male
Panel A: Days Absent								
Share Female Classmates	-10.00**	-5.27	-9.19**	-3.92	-10.30**	-5.46	-9.02**	-5.09
	(4.23)	(5.78)	(4.25)	(5.38)	(4.32)	(5.91)	(4.01)	(5.08)
Observations	2597	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. 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, Chernozhukov, and Hansen (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.

C Heterogeneity Results

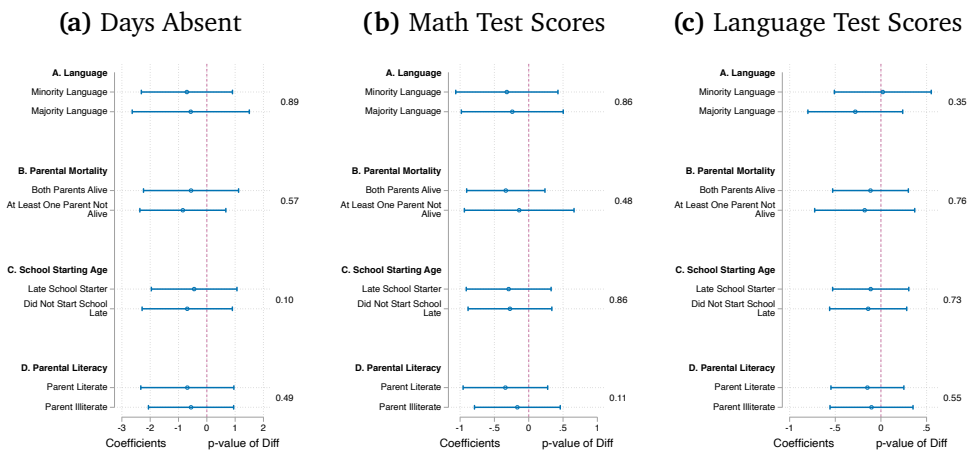
Figure C.1. Heterogeneity by Student Characteristics - Female Sample

(a) Days Absent **(b)** Math Test Scores **(c)** Language Test Scores



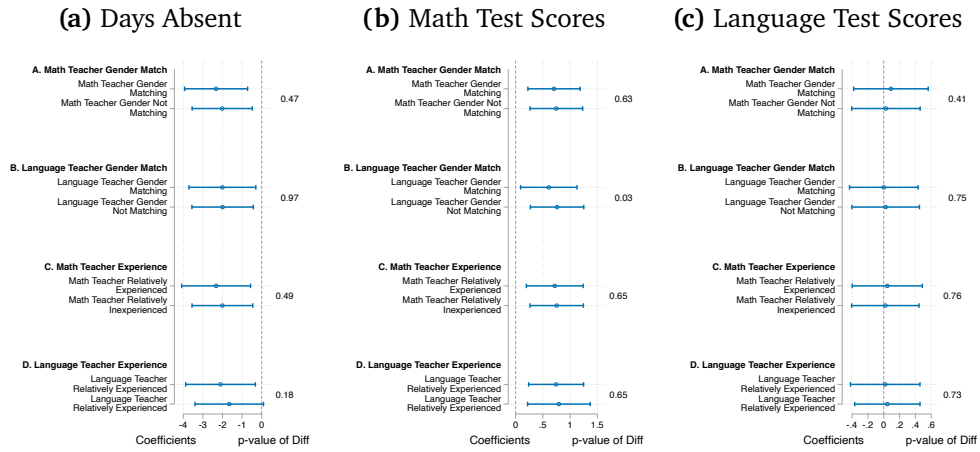
Notes:This figure presents heterogeneous effects of different subgroups on our outcomes including 95% confidence intervals clustered at the school level. We interact the share of female classmates variable with indicators of the respective student characteristics. The dependent variable is days absent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

Figure C.2. 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).

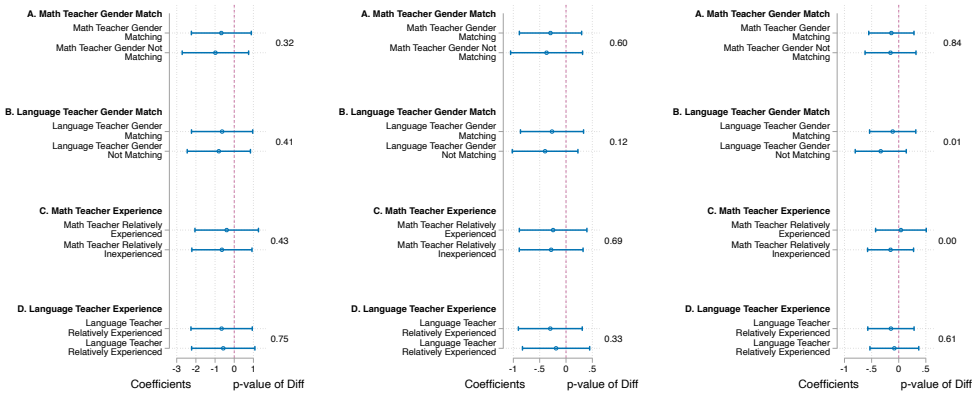
Figure C.3. Heterogeneity by Teacher Characteristics - Female Sample



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).

Figure C.4. Heterogeneity by Teacher Characteristics - Male Sample

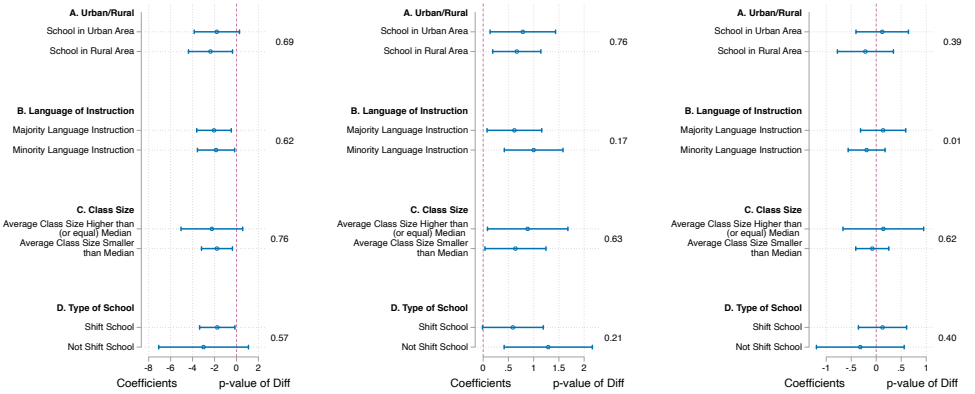
(a) Days Absent **(b)** Math Test Scores **(c)** Language Test Scores



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 C.5. Heterogeneity by School Characteristics - Female Sample

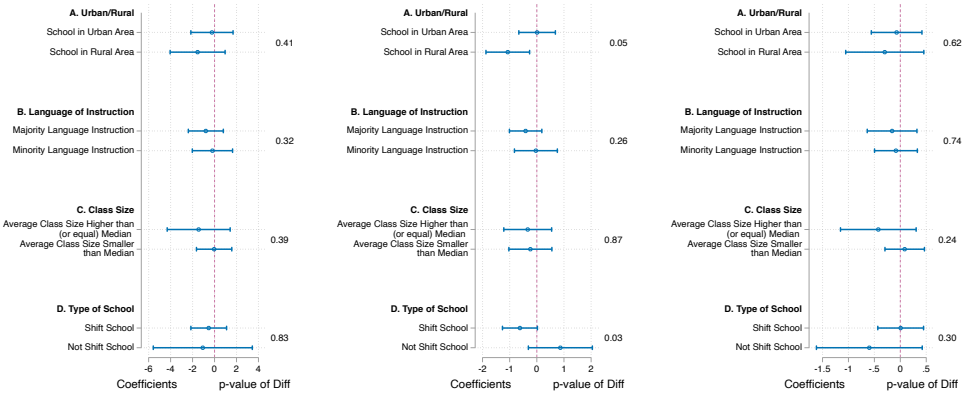
(a) Days Absent **(b)** Math Test Scores **(c)** Language Test Scores



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 C.6. Heterogeneity by School Characteristics - Male Sample

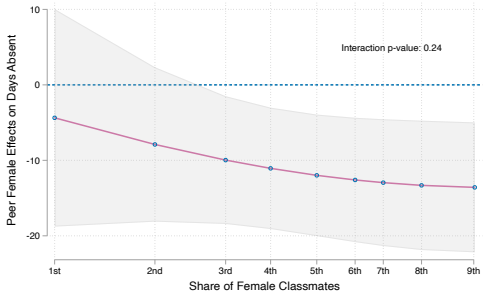
(a) Days Absent **(b)** Math Test Scores **(c)** Language Test Scores



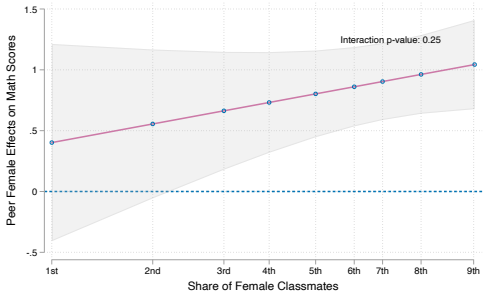
Notes:This figure presents heterogeneous effects of different subgroups on our outcomes including 95% confidence intervals clustered at the school level. We interact the share of female classmates variable with indicators of the respective school characteristics. The dependent variable is days absent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

Figure C.7. Nonlinearity in Classmate Gender Composition: Effects on Females

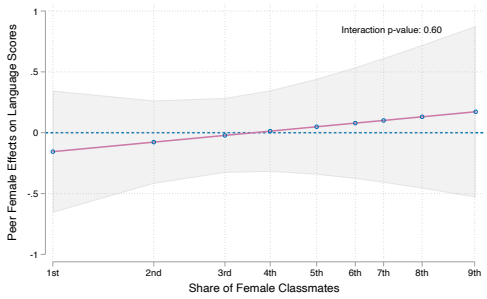
(a) Days Absent



(b) Math Test Scores



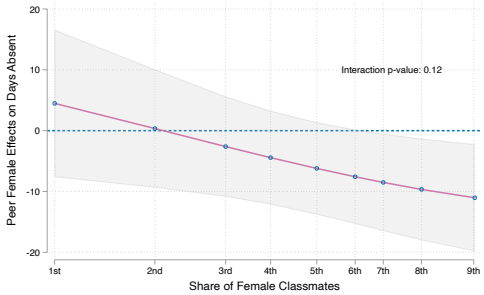
(c) Language Test Scores



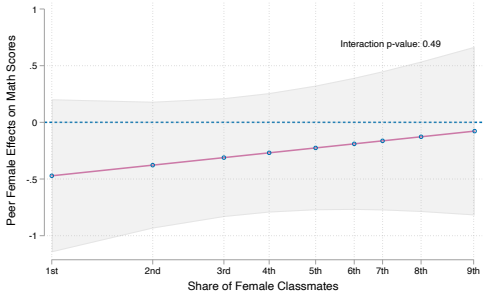
Notes: This figure presents the mean effects of the share of female classmates at deciles of peer female for females. It is based on our preferred baseline specification adding a quadratic in peer female on the subsample of females in the data. For days absent, we report the marginal effects based on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

Figure C.8. Nonlinearity in Classmate Gender Composition: Effects on Males

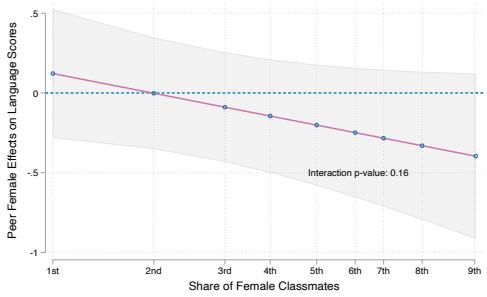
(a) Days Absent



(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.

D Additional Results for Mechanisms

Table D.1. Peer Gender Effects by Peer Age - Male Sample

	Days Absent		Math Scores		Language Scores	
Male Peers Age Tertile	Bottom Two	Oldest	Bottom Two	Oldest	Bottom Two	Oldest
Female Peers = Bottom Two	-1.40	6.77	0.09	-0.07	0.39	-0.42
	(4.82)	(13.40)	(0.41)	(0.35)	(0.27)	(0.34)
Female Peers = Oldest	-4.38	-3.03	-0.52	-0.60	-0.15	-0.64
	(12.13)	(9.60)	(0.65)	(0.58)	(0.61)	(0.52)
Observations	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.

Table D.2. Mechanisms - Bullying

	Bullied by Peers		
	(1) OLS	(2) OLS with Controls	(3) PDS Lasso
Panel A: Effect on Females / Males			
Share Female Classmates * Female	-0.41	-0.47	-0.40
	(0.42)	(0.44)	(0.39)
Share Female Classmates * Male	0.15	0.11	0.14
	(0.36)	(0.36)	(0.33)
Observations	490	489	489
# Unpenalized Controls			0
# Penalized Controls			22
# Selected Controls			5
Panel B: Mean Effect			
Share Female Classmates	-0.12	-0.17	-0.12
	(0.26)	(0.26)	(0.25)
Observations	490	489	489
# Unpenalized Controls			0
# Penalized Controls			22
# Selected Controls			2
Controls	Only Gender	Yes	Yes
School Fixed Effects	Yes	Yes	Yes

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 indicator for whether a student has even been bullied as the outcome variable. Column 1 uses only gender as a control variable. Column 2 uses the full set of our baseline control variables. Column 3 uses a post-double selection lasso model to select the control variables that are the best predictors of both the outcome and peer female composition.