Child beliefs, societal beliefs, and teacher-student identity match

Alex Eble
Teachers College,
Columbia University

Feng Hu
University of Science and
Technology Beijing

Children routinely benefit from being assigned a teacher who shares an identity with them, such as gender or ethnicity. We study how student beliefs impact teacher-student gender match effects, and how this varies across subjects with different societal beliefs about differential ability by gender. A simple model of belief formation predicts that match effects will be larger for students who believe they are of low ability, and be greater in subjects with more salient societal beliefs. We test these using data from Chinese middle schools, exploiting random assignment of students to teachers. In China, many people believe boys are innately better than girls at math. We find that being assigned a female math teacher helps low-perceived-ability girls and slightly harms low-perceived-ability boys, with no effects for other children. In English and Chinese – subjects with less salient societal beliefs – these patterns persist but diminish. This yields policy implications for the assignment of teachers to students.

VERSION: November 2019

Child beliefs, societal beliefs, and teacher-student identity match

Alex Eble and Feng Hu*

October 2019

Abstract

Children routinely benefit from being assigned a teacher who shares an identity with them, such as gender or ethnicity. We study how student beliefs impact teacher-student gender match effects, and how this varies across subjects with different societal beliefs about differential ability by gender. A simple model of belief formation predicts that match effects will be larger for students who believe they are of low ability, and be greater in subjects with more salient societal beliefs. We test these using data from Chinese middle schools, exploiting random assignment of students to teachers. In China, many people believe boys are innately better than girls at math. We find that being assigned a female math teacher helps low-perceived-ability girls and slightly harms low-perceived-ability boys, with no effects for other children. In English and Chinese – subjects with less salient societal beliefs – these patterns persist but diminish. This yields policy implications for the assignment of teachers to students.

*Eble (corresponding author): Teachers College, Columbia University. Email: eble@tc.columbia.edu. Postal address: Department of Education Policy and Social Analysis. 525 W. 120th St. New York, NY 10027. Hu: School of Economics and Management, University of Science and Technology Beijing. Email: feng3hu@gmail.com. Author names are shown alphabetically; both authors contributed equally to this research. This paper updates and replaces our previously circulated paper entitled "Stereotypes, Role Models, and the Formation of Beliefs." The authors are grateful to Anjali Adukia, Joe Cummins, John Friedman, Morgan Hardy, Asim Khwaja, Ilyana Kuziemko, Bentley MacCleod, Jordan Matsudaira, Derek Neal, Randy Reback, Jonah Rockoff, Judy Scott-Clayton, and Felipe Valencia for generous input, as well as seminar audiences at Chicago, Columbia, Fordham, Stanford, the 2017 IZA transatlantic meeting, NEUDC, and PacDev. Funding: This work was supported by the Humanities and Social Science Fund of Ministry of Education of China (grant no. 19YJA790029) and the National Natural Science Foundation of China (grant no. 71373002). Declarations of interest: none. Keywords: gender; beliefs; human capital; teacher-student gender match; behavioral economics. JEL codes: D83; I24; J16.

1 Introduction

When female or minority students are assigned to teachers of the same gender or race, they often learn more, perform better in school, and are more likely to pursue fields related to the subject taught by the teacher (c.f., Dee 2004; Bettinger and Long 2005; Dee 2005; Muralidharan and Sheth 2016; Lim and Meer Forthcoming). But why? Prior research has identified two main pathways. One pathway is that these teachers take actions that may lead to more effective teaching for these students, such as using different teaching methods, praising students of these groups more, or giving more opportunities for these students to participate (e.g., Hoffmann and Oreopoulos 2009; Lim and Meer 2017). The other pathway is that these teachers may affect student beliefs about the world, themselves, and their future, for example, by serving as role models or by having higher expectations for the child (e.g., Gershenson et al. 2016; Kofoed et al. 2017). There is still substantial uncertainty about how this second pathway works, for whom, and how this varies across subjects with different societal beliefs about girls' innate ability. In this paper, we attempt to address this gap by studying how student beliefs and societal beliefs impact teacher-student identity match effects.

First, we show conceptually how students' baseline beliefs about their own ability may make them more or less likely to respond to being assigned a teacher of the same gender. We also show how societal beliefs about differential ability by gender in certain subjects play a crucial role in this pathway. Children's beliefs about themselves and the world are influenced by their exposure to messages from society (e.g., Jayachandran 2015; Rodríguez-Planas and Nollenberger 2018). We focus on the belief that boys are inherently better than girls at learning math. Exposure to these societal beliefs distorts student beliefs about their own math ability (Nollenberger et al., 2016; Bordalo et al., 2019). Our conceptual framework predicts that, as a result of this distortion, low perceived ability students should be even more responsive to teacher-student gender match in math than in other subjects.

Second, we show empirically that evidence of these theoretical predictions appears in nationally representative data from China with random assignment of children to middle school teachers.

This empirical setting has two key features which facilitate our analysis. First, there is widespread belief among children and parents that boys have greater innate math ability than girls (Tsui, 2007). This occurs despite the fact that, on average, girls perform as well as or better than boys in mathematics. Second, there is random assignment of students to classrooms, which gives us random assignment of teacher gender to student gender.[1]

Our empirical results show that the existence and magnitude of teacher-student gender match effects vary systematically with child beliefs. As predicted, these effects are largest in math. Girls who believe they are of low ability in math and who are assigned to female math teachers gain relative to all other low perceived ability children. They are 20 percentage points less likely to perceive their current math class as "very difficult" (from a baseline of 80%), are 11 percentage points less likely to aspire to jobs in the visual or language arts (baseline 23%), and score 0.45 standard deviations (SD) better on a standardized math exam. When boys who believe they are of low ability in math are assigned to female math teachers, their outcomes appear to deteriorate relative to those of similar boys assigned to male math teachers: they are 10 percentage points *more* likely to perceive math as very difficult and experience a non-significant 0.15 SD drop in their math exam score. This is consistent with the idea that being assigned a female math teacher can counter the influence of societal beliefs about innate gender differences in math ability that may have inflated these boys' beliefs about their own ability.

We observe meaningful but smaller (and often insignificant) effects for low perceived ability girls assigned to female teachers in English and Chinese classes. This pattern is consistent with the fact that in these subjects there is no prevailing societal belief that girls' innate ability is inferior to boys'. Finally, we find very little evidence of teacher-student gender match effects on the beliefs, aspirations, or test scores of girls or boys who enter middle school believing themselves to be of average or high ability in math, Chinese, or English.

Our results are robust across three different classifications of low perceived ability, and they

[1] This source of plausibly exogenous variation has been used in several other studies of classroom configuration, both using data from Chinese middle schools (e.g., He et al. 2017; Gong et al. 2018) as well as in other countries (e.g., Lavy and Schlosser 2011; Lim and Meer 2017).

persist when a subset of the students are tested again a year later. We run a battery of tests for the possibility that variation in teaching methods, teacher aptitude, or teacher effort between male and female teachers drives the effects we observe. We find no evidence that variation in either the teaching methods used by the teacher or teacher aptitude differentially affect low perceived ability students in any of the three subjects. We find no evidence that female math teachers disproportionately call on or, separately, praise, low perceived ability girls relative to other students.

Our findings have clear implications for policy and research. For policy, in many settings there are relatively few female or minority teachers (c.f., Graham, 1987; Antecol et al., 2015). Our results suggest prioritizing the assignment of these teachers to children who share their identities and who perceive themselves to be of low ability in a given subject. This recommendation is even stronger when that child's belief corresponds to a social belief about differential ability by identity, e.g., girls are innately bad at math. For research, our findings suggest several avenues for further analysis. Do similar gains accrue differentially for other low perceived ability children who face societal beliefs about their ability, such as African American students in the US or students of non-caucasian descent in Europe and Latin America? If so, in what subjects are these effects greatest?

We advance the broad teacher-student identity match literature (c.f., Hoffmann and Oreopoulos, 2009; Carrell et al., 2010; Lim and Meer, 2017; Gershenson et al., 2018) by showing, both conceptually and empirically, that the power of shared-identity teachers to help students may depend crucially on two factors: students' own beliefs about their ability and the extent to which exposure to societal views (in this case, beliefs about differential math ability by gender) shape them. This finding complements recent work showing that shared-identity teachers serve as role models in other contexts, shaping career choices of college students who face similar societal beliefs in the US (Carrell et al., 2010; Kofoed et al., 2017; Porter and Serra, 2019).

We also provide new evidence on an interesting question: why does teacher-student identity match still matter in a setting like China, where girls perform better than boys in almost all subjects? In an increasing number of settings and subjects, there is no longer a gender gap in favor of boys (c.f., Ellison and Swanson, 2010; Lavy and Megalokonomou, 2019). Our paper explores this question in Chinese middle schools, where girls perform better than boys, but many parents and children still hold the traditional view that boys are innately better than girls at learning math (Tsui, 2007). Our results show that low-perceived-ability girls still benefit from having a female math teacher despite there being no overall gender gap in math performance. This complements other work documenting the persistent negative effects of these societal beliefs on the most vulnerable even when, on average, historically marginalized groups' outcomes have improved (Nollenberger et al., 2016; Rodríguez-Planas and Nollenberger, 2018).

The rest of the paper proceeds as follows: Section 2 presents our conceptual framework, Section 3 briefly describes the setting we study, Section 4 describes our data sources and empirical strategy, Section 5 presents our main empirical results, Section 6 tests alternative explanations for our results and discusses the probable mechanisms behind our results, and Section 7 concludes.

2 Conceptual framework

In this section, we generate the predictions that we test later in the paper. To do so, we use a simple model of bayesian updating to illustrate the potential for interaction between child beliefs about their own ability, societal beliefs about differential ability by gender in certain subjects, and teacher-student gender match. The key intuition is that the main effects of teacher-student gender match should accrue to children whose beliefs are more malleable to begin with, and that these effects are likely to be greatest in subjects where children's beliefs have been distorted by exposure to societal beliefs about differential ability by (gender) identity.

There are multiple ways that teacher-student identity match may affect children's learning. Using the framework of Dee (2005), these can be divided into "active" and "passive" effects. Active effects capture the extent to which same-identity teachers may behave differently than teachers of different identities and how this difference in behaviors may affect child outcomes. For example, these teachers may use different methods, language, or examples than those used by non-shared identity teachers. This may be more effective for shared-identity students than the methods used by teachers outside of the identity group. These teachers may also favor students of their own identity group with more praise or opportunities to express themselves (c.f., Lim and Meer, 2017).

Passive effects refer to the notion that shared-identity teachers may also affect students' beliefs and performance without any deliberate action on the teacher's part. These effects could work through the teacher's presence in the classroom and the messages it sends to children, or through the other messages the teacher may passively transmit to students about students' place in the world. For example, teachers are, by definition, experts on the subject matter they teach. Assigning students to a teacher with whom they share an identity, such as gender or ethnicity, may convey new information about the student's own potential to succeed in that subject. This is particularly the case for groups who may not have been exposed to other examples of this identity-specific potential for success in a given subject (c.f., Carrell et al., 2010; Wilson, 2012; Kofoed et al., 2017). The passive channel also includes the possibility that these teachers may have higher expectations for children with whom they share an identity, as in Gershenson et al. (2016). This also works to convince students that they may have better possible outcomes than previously thought.

In our model, we focus on these passive effects and the information they transmit to children. We begin with a child who believes with some probability P that they are of "high ability" in a given subject. We model all teachers as sending the message that each child can succeed in the subject being taught. Using Bayes' rule, we can model how children's beliefs about their own ability (P) respond to a given source of information (Camerer, 1995).

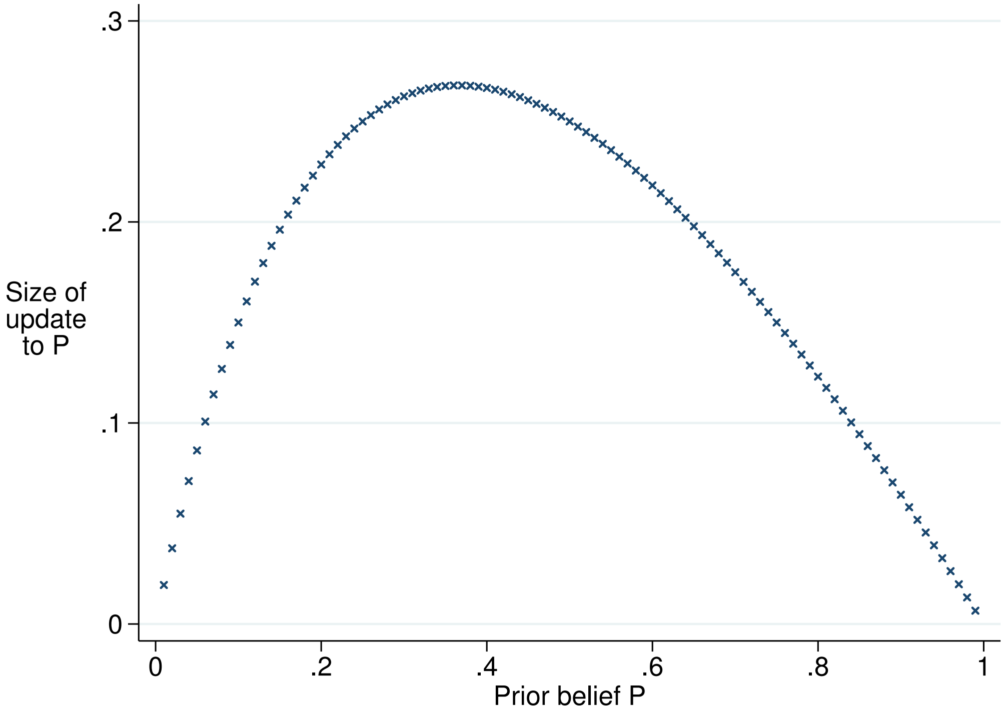
Our core assumption is that teachers who share an identity (e.g., gender) with the student send a more "credible" message about the student's potential for success in the subject, which causes a greater belief update. This comes from two strands of literature. First, the greater credibility and informational content of the messages sent by shared-identity role models, including teachers, is explored in several studies across economics and sociology (e.g., Bettinger and Long 2005; Hoffmann and Oreopoulos 2009; Wilson 2012). Second, a core tenet of information theory is that more credible messages cause greater updates to beliefs (e.g., Camerer 1995; Gärdenfors 2003;

Ambuehl and Li 2018).

The potential impacts of these messages also differ across two other child-level factors: the child's beliefs about their own ability and the societal beliefs they have been exposed to. Children can have a high or low P. Our first main result is that children with a low P are much more likely to update their beliefs than those with a high P.[2] In Figure 1, we show this relationship. On the X-axis, we plot a child's belief about P; on the Y-axis, we show how much child with a given P will update P in response to encountering the message, sent by the teacher, that they are of high ability. The hump-shaped figure shows the core insight: children who are uncertain about their P update their beliefs much more than children who are more certain.

[2] Children who have performed consistently poorly will have a very low P. Because the demand for education is so high in China, overall performance is quite strong, and few students in in our data perform this poorly, we abstract away from such cases of "certainty" about low ability here.

Figure 1: How prior beliefs about ability map onto to the size of an update in beliefs



Note: This figure shows the mapping from P, a child's prior belief about the likelihood that they are of high ability, to the update of that belief in response to a new source of information telling the child they are likely to be of high ability (for example, encountering a same-gendered teacher). We generate this data by with the following assumptions: $Pr(HighAbility|TeacherStudentGenderMatch) = 0.6$, i.e., the perceived probability of encountering a same-gendered teacher, given that you are of high ability, is 0.6, and $Pr(HighAbility|NoTeacherStudentGenderMatch) = 0.1$. As long as the perceived probability of being high ability is greater with match than without it, the hump shape of the distribution holds under all but a trivially small set of values of these two variables.

Next, we introduce the impacts of societal beliefs. We study the impacts of exposure to the societal belief that boys are inherently better than girls at learning math. Exposure to this belief has been shown to distort children's beliefs about themselves, depressing girl's beliefs about their own math ability and inflating boys' (Nollenberger et al., 2016; Bian et al., 2017; Rodríguez-Planas and Nollenberger, 2018; Eble and Hu, 2019). These beliefs directly contribute to worse performance among girls via two channels. First, anxiety because of "stereotype threat" (Shih et al., 1999; Spencer et al., 1999; Niederle and Vesterlund, 2010; Cheryan, 2012) may reduce performance on math assessments. Second, negative gender norms may exert downward pressure on a child's beliefs about their returns to investment, causing girls to invest less effort, enthusiasm, and time in studying for math (Bian et al., 2017).

In our framework, exposure to the societal belief that boys are better than girls at learning math introduces systematic bias into the beliefs of children. It artificially inflates boys' beliefs about their math ability, and artificially suppresses girls' beliefs about their math ability. The impacts of societal beliefs will again be the strongest for children who already, for whatever reason, have lower P, as their beliefs are more malleable.

Encountering a female *math* teacher - as opposed to a female English or Chinese teacher - therefore provides a particularly strong signal to both girls and boys. This is because the presence of a woman teaching their math class runs directly counter to this societal belief. This exposure should increase girls' perception of the likelihood they may be able to succeed in math, while providing evidence to boys that they are not necessarily good at math just because of their gender.

To summarize, our framework generates two core predictions for teacher-student gender match effects. The first prediction is that the incidence and magnitude of these effects should vary with children's baseline beliefs about their own ability in the subject. The second prediction is that they should also depend on the extent to which there exist societal beliefs about ability by identity (gender) in that same subject. Note that for this second prediction, there are two gender-specific predictions. In addition to the anticipated salutary effects for girls, we predict that, for boys who perceive themselves to be of low ability in math, being assigned to a female teacher in a new, more challenging math class may threaten their perceptions that their math ability is inherently superior to girls'.

3 Setting

In our empirical work, we study a nationally representative sample of students who entered Chinese middle schools in the early 2010's. Since 1986, middle school has been compulsory in China.

Primary school graduates are assigned to their neighborhood middle school following the local educational authority's districting. Because a child's school is determined by place of residence and families are only allowed to send their children to schools in the area where their household residence permit was issued, there is little scope for sorting into schools/school districts with(out) random assignment. Furthermore, as we describe later, our empirical strategy uses variation between classes within each school, minimizing the potential threats to internal validity from any remaining sorting across schools.

Since 2006, a separate law has banned tracking of students to different classes based on demonstrated ability or academic performance. As a result, there are currently two permitted methods of assigning students to classes in China's middle schools: (1) purely random assignment and (2) assignment of students to maintain similar average levels of performance across classes, based either on students' academic performance on primary school graduation examinations or on diagnostic examinations arranged by the middle school. In the first system, students are then randomly assigned to classes by lottery or another quasi-random method.[3] In the second system, students are assigned to classes by an algorithm which takes into account their academic performance at the beginning of the seventh grade and enforces a "balanced assignment" rule, requiring that the average quality of students be comparable across classes and the class not be bifurcated (Carman and Zhang, 2012).[4] Several recent papers exploit this random assignment of students to classes and provide explanations of the two different assignment mechanisms (Hu, 2015; He et al., 2017; Gong et al., 2018).

We follow prior work studying this setting, exploiting these two methods of assigning students to classes as providing potentially quasi-random matching of student gender to teacher gender (c.f., He et al., 2017; Gong et al., 2018). As described in Hu (2015) and Gong et al. (2018), who use the same data as we do, this system is not implemented with perfect fidelity. Despite the banning of class tracking, as students progress through middle school some schools may assign students to classes based on their academic performance in order to better prepare top students for the entrance

[3] For instance, according to alphabetical order by surname, i.e., every n^{th} student assigned to the n^{th} class.

[4] In Appendix C, we provide a description of this type of assignment rule, borrowing from He et al. (2017).

examination; this practice is more common in the eighth and ninth grades than in the seventh. In our analysis here, as in those three papers, we restrict our attention to students randomly assigned to classes in the seventh grade and to students in the ninth grade in those schools where random assignment of students to classes is maintained throughout middle school.

We build on the work of two papers which use this same data to study teacher-student gender match effects. Gong et al. (2018) use the first round of the same data we use to study the salutary effects of teacher-student gender match on *all* female students' academic outcomes and non-cognitive skills. Xu and Li (2018) focus on the persistence of mathematics effects across both waves. We study how child and societal beliefs determine the existence and magnitude of these teacher-student gender match effects. Our work illuminates a core mechanism behind the empirical results they find and suggests potential avenues for better understanding teacher-student gender match effects in other contexts.

4 Data and empirical strategy

This section describes our data sources and empirical approach. Section 4.1 outlines the data we use and provides summary statistics. Section 4.2 describes how we classify students as "low perceived ability." Section 4.3 describes the empirical strategy we use, and Section 4.4 states and tests our main identifying assumptions.

4.1 Data sources and descriptives

The main data source we use in this paper is the China Education Panel Survey (CEPS) conducted by the National Survey Research Center at Renmin University of China. The CEPS is a nationally representative longitudinal survey that aims to track middle school students through their educational progress and later labor market activities. Its sample was selected using a stratified, multistage sampling design with probability proportional to size, randomly selecting approximately 20,000 seventh and ninth grade students from 438 classes in 112 schools from 28 counties across mainland China during the 2013-2014 academic year. In each selected school, four classes were randomly chosen, two from the seventh grade and two from the ninth. All students in the selected classes were then surveyed. The CEPS uses five different questionnaires, administered to students, parents, homeroom (banzhuren) teachers, main subject (math, Chinese, and English) teachers, and school administrators, respectively. It is China's first nationally representative survey targeting middle school students, comparable to the Adolescent Health Longitudinal Studies (AddHealth) in the U.S. and the National Education Panel Survey (NEPS) in Europe.

The CEPS contains rich demographic data on students, their families, and their teachers, as well as detailed information on students' beliefs, aspirations, and time use. It also collects administrative school records on students' midterm test scores in three compulsory subjects: math, Chinese, and English. The scores are standardized in terms of school and grade, with a mean of 70 and a standard deviation of 10. They are (relatively) low stakes exams, graded collectively by the math teachers in the student's grade. Although their grading is not always blinded, Gong et al. (2018) argue that blinded grading is common in these particular tests. We make a slightly weaker argument: that these low stakes exam scores are unlikely to be substantially biased by teacher gender, and it is even less likely that they will be differentially biased for low perceived ability girls assigned to female teachers.

The survey also collects data on the assignment mechanism used to assign students to classrooms, collected both from school principals and homeroom teachers.[5] The options are 1) tracking; 2) assignment according to students' household registration location; 3) either literally random assignment ("sui ji", meaning random) or assignment according to the average-equilibrating algorithm described above; or 4) assignment through other methods. About 85% of middle schools in our data assigned entering students to classes in either a random or an average-equalizing manner.

[5] This data is self-reported. We argue that reporting bias in the assignment mechanism data is unlikely because the data collection process stresses the anonymity of the data (all identifying information is removed from the datasets released to scholars) and the data is collected by academics and graduate students, not government officials. We also limit the analysis to grades where both school principals and homeroom teachers report use of random assignment. Homeroom teachers are less likely than principals to face potential negative consequences of the school using a non-random assignment mechanism, and this restriction serves as a further check on the principal's self-report. Hu (2015), Gong et al. (2018), and Xu and Li (2018) use these same data points and also describe their reliability for determining assignment mechanism.

Among those schools, one third reassigned students based on past academic performance when they entered the eighth or ninth grade.

We use the same sample restriction as used in Hu (2015) and Gong et al. (2018), which treats assignment to class as random for two sets of children. The first set is seventh graders in schools which report the use of either purely random assignment or the average-equalization algorithm to assign seventh-grade students to classes. The second set is ninth graders in the subset of these schools which also report not reassigning eighth and ninth grade students to new classes in terms of previous academic performance after initial quasi-random assignment in the seventh grade. Both Hu (2015) and Gong et al. (2018) also demonstrate the validity of this approach for causal inference using these data.

Table 1 presents summary statistics for students by gender for those students randomly assigned to classrooms. Among the children in our sample, the average age of girls is lower than that of boys, and girls are more likely to have more educated parents and higher family incomes. Girls in our sample also have more siblings than boys, a consequence of the prevailing son-favoring tradition and the birth control policy in China, which allows for multiple children in some cases if the first child is a girl. Finally, girls perform better than boys on math, Chinese, and English tests administered at the school level.

Table A.1 shows summary statistics for teachers in the classrooms studied in Table 1. Thirty-nine percent of the math teachers in our data are male, alleviating the challenge faced in Antecol et al. (2015) where the small number of male teachers restricted the strength of the conclusions they could draw from their estimates of the effects of teacher-student gender match; 20 percent of Chinese teachers are male, and nine percent of English teachers are male. Female teachers are on average younger and less experienced than their male counterparts. Overall, female teachers appear to be slightly more qualified than their male counterparts in terms of education and proportion having won a teaching award at the province or national level.[6] The observed differences described here attenuate dramatically and cease to be significant at the level of comparison we study in

[6] A teaching award at the national level is the most prestigious, followed by an award at the province level, and awards at the city level (the smallest of the three geographical units) are the least prestigious.

Table 1: Summary statistics for students

	(1) All children	(2) Girls only	(3) Boys only	(4) Difference (column 2 - (column 3)	(5) P-value of difference
Proportion female	0.480	—	—	—	—
Age	13.23	13.17	13.28	-0.11	0.00
Holds agricultural hukou	0.49	0.48	0.51	-0.03	0.04
Number of siblings	0.71	0.76	0.66	0.10	0.00
Household is poor	0.19	0.18	0.20	-0.02	0.01
Father's highest credential					
Middle school	0.41	0.41	0.42	-0.01	0.44
High school	0.26	0.25	0.26	-0.01	0.69
College	0.19	0.20	0.18	0.02	0.03
Mother's highest credential					
Middle school	0.38	0.39	0.37	0.02	0.01
High school	0.23	0.23	0.23	0.00	0.56
College	0.16	0.17	0.16	0.01	0.12
Ethnic minority	0.12	0.12	0.11	0.01	0.23
Math test score	70.1	70.9	69.4	1.50	0.00
English test score	70.1	73.0	67.4	5.60	0.00
Chinese test score	70.0	73.2	67.1	6.10	0.00
Number of observations	9,361	4,492	4,869	—	—

Note: This table describes student characteristics overall (column 1), and then separately by gender (columns 2 and 3). Column 4 tabulates the difference in means for each characteristic between girls and boys, and column 5 shows the p-value of a test for whether any of these gender differences are statistically significant. It uses only data from the main estimation sample in the paper, described in Section 4.1.

our paper: comparing between children within a school, within a grade, between children in a classroom with a female teacher and those in a classroom with a male teacher.

4.2 Classifying students as “low perceived ability”

We focus on two different specifications for classifying students as low perceived ability in a given subject. The CEPS asks students how difficult they found learning math, Chinese, and English (respectively) in the sixth and final grade of primary school. We use this to generate our two low perceived ability classifications. Specifically, we classify those students who found learning a given subject in the sixth grade to be “very difficult” as low perceived ability in that subject, and classify those who report that they found the sixth grade class in question to be “somewhat difficult,” “not so difficult,” or “easy” as not of low perceived ability in that subject. Note that this is not intended to proxy for a student’s actual ability, but rather, as a (noisy) measure of how able they perceive themself to be.

In Table A.2 we show characteristics of students, by gender, for both of the perceived ability groups, separately for math, Chinese, and English. The performance gaps between boys and girls described earlier appear here. Nonetheless, consistent with the societal beliefs we study, a higher proportion of girls perceive themselves to be of low perceived ability in math than do boys (11.7% vs. 8.9%). In our alternative specification, we classify respondents who report finding the subject to have been either “very difficult” or “somewhat difficult” in the sixth grade as low perceived ability.

To further test the robustness of our results, we also conduct a third, parallel analysis. This analysis looks separately at the performance of students whose scores fall below the median within their own teacher-student gender pairing type (girls assigned to a female teacher; girls assigned a male teacher; boys assigned a female teacher; and boys assigned a male teacher) in a given subject. Our main findings are robust to choice of specification.

4.3 Empirical strategy

In this subsection we explain our empirical strategy. We estimate the effects of being assigned a female teacher on female and on male students' outcomes and how these effects vary by students' perceived ability in each subject. We estimate the following reduced form regression equation:

$$\begin{split} Y_{icgj} = \beta_0 + \beta_1 FS_{icgj} + \beta_2 FT_{cgj} + \beta_3 (FS_{icgj} * FT_{cgj}) + \gamma_0 LPA_{icgj} + \gamma_1 (LPA_{icgj} * FS_{icgj}) + \\ \gamma_2 (LPA_{icgj} * FT_{cgj}) + \gamma_3 [LPA_{icgj} * (FS_{icgj} * FT_{cgj})] + \beta_4 SC_{icgj} + \beta_5 TC_{cgj} + \eta_{gj} + \varepsilon_{icgj} \end{split} \tag{1}$$

This specification follows other recent work studying teacher-student gender match, e.g., Muralidharan and Sheth (2016), and Lim and Meer (2017). The variables are defined as follows: Y_{icgj} denotes the outcome of interest for student i in class c of grade g in school j. FS_{icgj} is an indicator equal to one if student i is female, and FT_{cgj} is also an indicator, equal to one if the subject teacher in class c in grade g of school j is female. LPA_{icgj} is an indicator equal to one if the student perceives themself to be of low ability in the subject. SC_{icgj} is a vector of predetermined characteristics at the student level, TC_{cgj} is a similar vector for teachers, η_{gj} is a set of grade-by-school fixed effects, and ε_{icgj} is a robust standard error, clustered at the school level to allow for heteroskedasticity and arbitrary serial correlation across students within a given school.[7]

We estimate this equation separately for each subject: math, Chinese, and English. Unless otherwise specified, the controlled-for student-level characteristics determined prior to assignment of teacher gender include age, ethnicity (either Han or non-Han), hukou status (agricultural or not), parents' education levels, the child's number of siblings, and a categorical measure of household income (low income or not). The teacher-level predetermined characteristics include teacher age, education level, years of work experience, whether the teacher graduated from a normal (i.e., teacher training) university, whether the teacher holds a senior rank within the school, and whether they have won teaching awards at the city, province, or national level, respectively.

Intuitively, our estimation strategy compares the academic performance of students who study in the same grade in a middle school and share background characteristics, but are randomly as-

[7] Clustering at the (less conservative) classroom level improves the precision of our results.

signed to either a female or male teacher. Our identifying assumption is that, by virtue of random assignment, the match of FS_{icgj} to FT_{cgj} is orthogonal to predetermined characteristics which may influence student beliefs or academic performance. We test this assumption later in this section.

All of our estimated coefficients display children's performance relative to non-low perceived ability boys assigned to a male teacher (the omitted category). The coefficients β_1, β_2, and β_3 indicate how all children with a certain characteristic (e.g., β_1: girls; β_2: children assigned to a female teacher; β_3: girls assigned to a female teacher) compare to this group. The coefficients γ_1, γ_2, and γ_3 indicate how low perceived ability children with these same characteristics (girls, students assigned to a female teacher, and the interaction) fare relative to low perceived ability boys assigned to male teachers.

To emphasize how this approach advances the teacher-student identity match literature, we present our initial results for math - for beliefs about ability in math and for actual academic performance in math - sequentially. First, we show results estimated using the standard teacher-student gender match specification, i.e., without the low perceived ability interaction terms, as is done in most prior work (e.g., Muralidharan and Sheth, 2016; Lim and Meer, 2017). Second, we present results from the fully specified model, which includes the low perceived ability variable and its interactions. For the sake of (relative) brevity, we show estimates from only the fully specified model for subsequent analyses.

In our exposition, we focus on three sets of estimates. The first set comprises β_3 in the standard model and β_3 and γ_3 in the fully specified model, which we interpret as quasi-experimental estimates of the effect of being assigned a female teacher on (low-perceived ability) girls relative to the effect for (low perceived ability) boys. This captures the effect of teacher-student gender match on girls' performance, relative to boys (Dee, 2007; Muralidharan and Sundararaman, 2011; Lim and Meer, 2017). In line with previous work on teacher-student identity match, these coefficients should be non-zero and point in the direction of reducing girls' perceived difficulty of the subject and improving their performance. The second set is γ_3 and β_3 in the fully specified model. Comparing the magnitude of γ_3 and β_3 in this specification is a test of the prediction that there will be larger effects for low perceived ability girls. The third set is γ_2 and γ_3. The coefficient γ_2 captures the effect of being paired with a female teacher on low perceived ability boys, using those low perceived ability boys assigned to a male teacher as the comparison group. We predict that γ_2 and γ_3 (i.e., the effects for low perceived ability boys and girls, respectively) will differ in sign.

There are several parameters of ancillary interest that are derived from different combinations of the coefficients we estimate in equation 1, and we will explicitly address a few of these in our discussion of the empirical results. First, $\gamma_2 + \gamma_3$ yields the total effect of being assigned a female teacher on low perceived ability girls, relative to low perceived ability girls assigned a male teacher. It is the sum of the effect of being assigned a female teacher on low perceived ability students and the effect of being assigned a female teacher specific to low perceived ability girls. At the bottom of our main results tables we show this parameter and the p-value for a test of whether it is statistically significant. Second, $\beta_3 + \gamma_3$ yields the total effect of teacher-student gender match on the gender gap for low perceived ability girls, i.e., making the comparison group all boys, not only low perceived ability boys.

4.4 Identification

If our assumption of orthogonality is satisfied, estimating equation 1 using OLS should recover unbiased estimates of these parameters. Several earlier papers using this data - for example, Hu (2015), Gong et al. (2018), and Xu and Li (2018) - investigate this claim. They show that, using the same sample restriction that we use, there is strong evidence of orthogonality. To further test the assumption that, within a grade within a given school, the match of student gender to teacher gender is as good as random, we follow the method of Hansen and Bowers (2008), Bruhn and McKenzie (2009), and Antecol et al. (2015).[8] We regress teacher gender on the same set of observable, predetermined student and family characteristics and grade-by-school fixed effects described above (that we include in our main empirical specification). For each regression we present coefficient

[8] Note that because we do not have baseline performance data, we cannot perform the synthetic classroom randomization test used in Carrell and West (2010) and Kofoed et al. (2017).

estimates and report the F-statistic and p-value from a Wald Test of the joint significance of the regressors. We present these results in Table 2. For each subject, our F-test fails to reject the null that the regressors are together not significant predictors of teacher gender (column 2). Though one of the twelve individual coefficients is statistically significant, this is consistent with statistical chance. In Appendix Table A.3, we also present the conventional regression of teacher gender on each controlled-for predetermined variable at the student level. Its results largely mirror those of 2, with only one of the estimates for math being statistically significant, and that one only at the 10% level (two are significant for Chinese, and three for English). These results support our main identifying assumption that students' observable predetermined background characteristics are balanced along the gender of subject teachers within the same grade in a given school.[9] While we cannot rule out the possibility that in some cases influential parents or individuals successfully lobbied to be placed with a certain teacher, we conclude from these results that such non-random matching of teachers to children is unlikely to be common enough to substantially bias our estimates.

Another descriptive comparison of interest is teacher quality across genders. To ensure that we are isolating the effect of teacher gender from other teacher characteristics, we need to establish whether male and female teachers differ on observable characteristics, such as teaching skill, which could drive any effects we measure (Cho, 2012; Antecol et al., 2015). To do so, we conduct an empirical test similar to that in Table 2, only with the analysis at the teacher level. The predetermined characteristics we include on the right hand side are age, a dummy for having earned a full-time bachelor's degree or higher qualification, a dummy for having attended a "normal" university (i.e., a university specializing in teacher training), years of teaching experience, and two dummies for having won a teaching award at two different levels, respectively. Our results, shown in Table A.4, fail to reject the null that within a grade within a school, these characteristics are not jointly predictive of the teacher's gender in any of the three subjects.

Note that if schools deliberately misreport their assignment mechanism - i.e., claiming random

[9] Though we would like to conduct a synthetic randomization test, as in Carrell and West (2010) and Kofoed et al. (2017), we lack pre-assignment performance data. As a result, we cannot further test our assumption that class assignment is orthogonal to student aptitude.

assignment when in fact they use tracking - this would bias upward our estimates of the effect of female teachers on the best students (i.e., β_3) and bias downward the effect on worse students (γ_3), who are less likely to be assigned to "good" teachers. This is because most school administrators are seeking to maximize the performance of the best students (Kipnis, 2011). In short, any bias from this type of misreporting would push our coefficient estimates in the opposite direction of our main predictions.

We note that our perceived ability data is observed at the same time as all of the other data, specifically, after teacher assignment. It is possible, therefore, that teacher gender could affect a child's report of the difficulty they had in a given primary school subject, possibly in a way that is correlated with controlled-for predetermined characteristics such as gender. We test for this possibility by running the same regressions of teacher gender on our list of predetermined characteristic controls, only restricting our analysis to low perceived ability students in each subject

We show our results in Table A.5. We find no evidence that predetermined student characteristics impact a child's likelihood of reporting low perceived ability (i.e., presence in the low perceived ability sample) in a way that is correlated with the gender of their teacher in any of the three subjects. The general pattern is the same as that for the entire sample in Table 2 - after controlling for grade-by-school fixed effects, only one of the 12 estimated coefficients is statistically significant and we fail to reject the null that these characteristics are jointly insignificant predictors of teacher gender.

5 Main empirical results

In this section, we present our main empirical results. We present them sequentially, starting with results for math then presenting results for all three subjects together.[10] For the math results, we

[10] There are multiple reasons for this order of presentation. First, our model generates ten separate coefficients of interest, and starting with one subject ensures clarity of exposition. Second, we wish to highlight our results for math, as it is the subject in which the predictions from our framework are the clearest. Third, we have greater statistical power for the analysis in math than for either English or Chinese. Specifically, in math, we have a good balance of female and male math teachers (60 and 40 percent, respectively) and several hundred low perceived ability children of both genders. In English, unlike in math and Chinese, there are relatively few male teachers (19 out of 210). In

Table 2: Test for random assignment

	(1) Math	(2) Chinese	(3) English
Number of siblings	-0.006	0.004	0.003*
	(0.006)	(0.005)	(0.002)
Household is poor	0.007	0.000	-0.003
	(0.013)	(0.011)	(0.004)
Female	0.002	-0.002	-0.001
	(0.005)	(0.003)	(0.003)
Age	-0.011**	0.001	0.007*
	(0.005)	(0.004)	(0.004)
Ethnic minority	0.012	-0.022	0.005
	(0.019)	(0.015)	(0.004)
Holds agricultural hukou	-0.010	0.013	0.004
	(0.013)	(0.009)	(0.005)
Mother's education level			
Middle school	0.009	0.001	-0.001
	(0.013)	(0.008)	(0.005)
High/technical school	0.002	-0.007	0.002
	(0.013)	(0.008)	(0.006)
College or above	0.003	0.000	0.014
	(0.016)	(0.008)	(0.011)
Father's education level			
Middle school	-0.012	-0.001	0.006
	(0.010)	(0.006)	(0.005)
High/technical school	-0.001	0.001	0.008
	(0.014)	(0.009)	(0.005)
College or above	0.008	0.005	0.000
	(0.017)	(0.012)	(0.007)
Low perceived ability in subject	-0.015	-0.005	-0.002
	(0.019)	(0.014)	(0.005)
Number of observations	8,155	8,085	7,910
R-squared	0.65	0.85	0.72
Joint test F-statistic	1.05	0.91	0.54
[p-value]	[0.42]	[0.55]	[0.91]

Notes: This table shows results from three separate omnibus regressions of teacher gender (=1 if female in the subject listed in the column heading) on the set of independent variables listed in the first column along with grade-by-school fixed effects as in our main estimating equation. This follows Bruhn and McKenzie (2009) and Antecol et al. (2015). We show the coefficients for the variable with the robust standard errors below. In Table A.3, we show one-by-one tests for orthogonality of teacher gender and predetermined characteristic. The results are similar.

present the usual teacher-student gender match specification - that is, without the low perceived ability control and its interactions - alongside our "full" specification, which does include these additional controls. Showing these side-by-side illustrates how our model advances understanding of teacher-student gender match effects in this setting. When we present analysis of math, Chinese, and English side-by-side, we present only results from estimating the full specification. For these, we present results using both the main and alternative definitions of low perceived ability. This shows, for each subject, how robust the patterns are to choice of classification.

5.1 Results for math

In this section, we present our results for teacher-student gender match in math. We first show analyses for the effects of match on students' beliefs about their own ability in math and their career aspirations. We then show results for midterm math test scores in year 1. We conclude the subsection with analysis of beliefs and test scores in year 2 for the subset of students interviewed in the second wave of the CEPS.

5.1.1 Beliefs about own math ability and aspirations

We first estimate the impact of teacher gender on two variables: perceived difficulty of the current math class and the careers to which students aspire. Our specification follows equation 1, using grade-by-school fixed effects and the full battery of controls for students and teachers.

For the analysis of perceived difficulty, we use students' response to the prompt "how difficult do you find your *current* math class to be?"[11] The potential responses are "very difficult," "somewhat difficult," "not so difficult," and "not difficult at all." We code this variable the same way we code the perceived ability variable: as an indicator equal to one if the response is "very difficult."

To study the impact of teacher-student gender match on aspirations, we use the child's response

Chinese, unlike in the other two subjects, there are very few girls who perceive themselves to be of low ability in Chinese using the primary definition (93 out of 4,463; roughly two percent, compared with nine to 15 percent for the other subjects).

[11] Recall that the baseline perceived ability question asked about the child's experience in the sixth grade; this question refers to the child's current experience in either the seventh or ninth grade.

to the prompt "what job would you most like to do when you grow up?" There are several possible responses to the question and we investigated two potential outcomes.[12] The first, on the lower end of aspirational change, is an indicator for whether or not the child aspired to jobs traditionally associated with women. In the raw data, women are most likely to choose jobs in the language and visual arts (designer; artist/actor), and we generate a variable equal to one if the job aspired to is one of these and equal to zero otherwise. The second variable indicates whether (1) or not (0) the child reports aspiring to be either a scientist or engineer.

We present these results in Table 3. We show coefficient estimates for the following parameters in our main estimating equation: for the "standard" model, we report the coefficients for whether the child is female (β_1), whether the teacher is female (β_2), and their interaction (β_3); for the fully specified model we add coefficient estimates on whether or not the child perceives themself to have low ability in math (γ_0), and the interactions between this variable and the three main variables from the standard model (γ_1, γ_2, and γ_3, respectively).

For each outcome, we first show results with the standard teacher-student gender match specification first, i.e., without any reference to low perceived ability children. We then present results from the fully specified model, which includes the low perceived ability variable and its interactions. For these results, we report the main coefficients ($\beta_1, \beta_2, \beta_3, \gamma_0, \gamma_1, \gamma_2, \gamma_3$) along with three additional parameters at the bottom of the table: the overall effect on low perceived ability girls ($\gamma_2 + \gamma_3$), the p-value for a statistical test of the null hypothesis that this parameter is equal to zero, and a yes/no description of whether or not we reject the null that the effect for low perceived ability girls ($\gamma_2 + \gamma_3$) is the same as for boys (γ_2).[13]

In column one of Table 3, we see the standard teacher-student gender match result: girls assigned a female math teacher (β_3) are nine percentage points less likely to find math difficult than girls assigned a male math teacher. In column two, the full specification shows that low perceived ability students are most affected by this match. Being taught by a female math teacher reduces

[12] The options are 1. Government Official, 2. Business manager, 3. Scientist/engineer, 4. Teacher/doctor/lawyer, 5. Designer, 6. Artist/actor, 7. Athlete, 8. Skilled worker, 9. Other, 10. Don't care, 11. Don't know.

[13] This is shown as "yes" when the coefficient estimate for γ_3 is statistically different from zero, and "no" when it is not. We present this yes/no classification separately for ease of exposition.

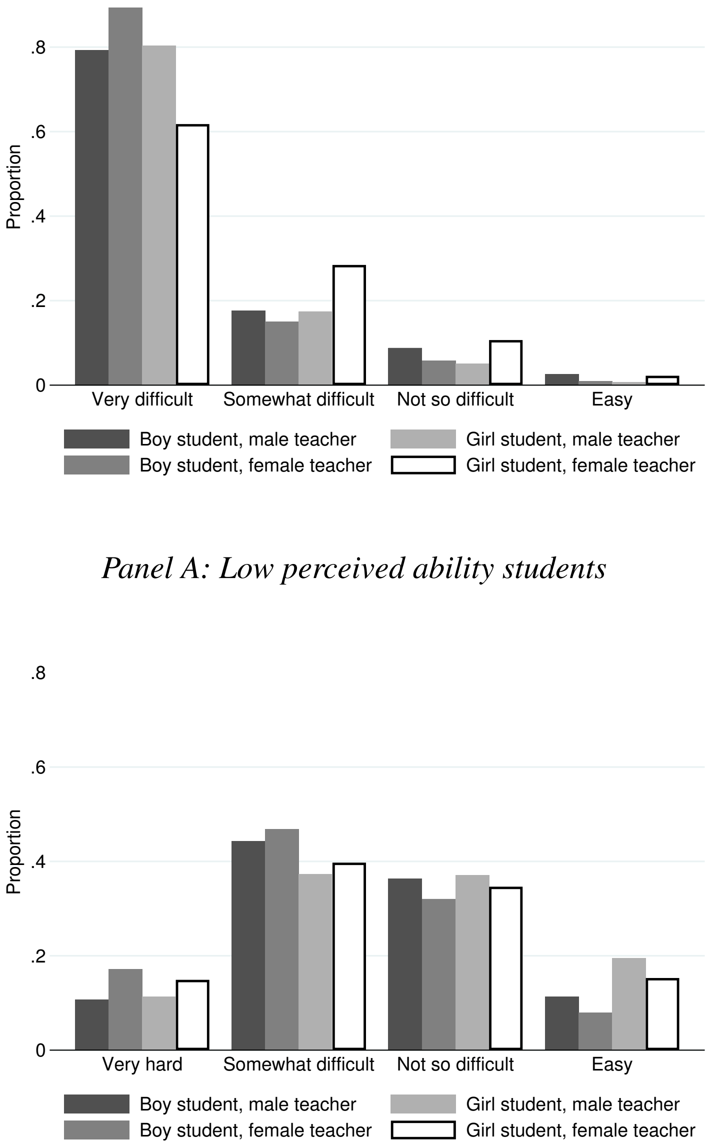
low perceived ability girls' probability of perceiving math as "very difficult" (γ_3) by nearly 23 percentage points. While the estimated effect for non-low perceived ability girls assigned to a female teacher (β_3) retains its sign, its magnitude diminishes by more than half. This new estimate of β_3 is nearly an order of magnitude smaller than that for γ_3. For low perceived ability boys, being assigned a female math teacher is associated with a 10 percentage point increase in the likelihood of reporting their current math class to be very hard (γ_2).

In Figure 2 we show two plots of the distribution of the perceived difficulty variable, one for low perceived ability children (Panel A) and another for the rest of the children in our sample (Panel B). The patterns in Panel A are very similar to the regression coefficients for low perceived children: low perceived ability ability girls assigned to a female math teacher are at least 20 percentage points less likely to perceive math to be very difficult than any other group, and low perceived ability boys assigned to female math teachers are at least 10 percentage points more likely to find math very difficult than any other group. Panel B shows no detectable difference in the perceived difficulty of the current math class between non-low perceived ability girls assigned to female teachers and all other groups.

In Table A.6 we show analog results for the alternative specification of low perceived ability. We observe an 11 percentage point decrease in the perceived difficulty of mathematics for low perceived ability girls thus defined. In column 1 of Table A.6, we present estimates generated using students below the within-group median test score instead of using either of the other low perceived ability group definitions. We observe that below-median girls assigned to a female teacher are 7.8 percentage points less likely to find math very difficult. These differences are smaller than what we find using the original definition of low perceived ability, but they retain both their sign and statistical significance.

Next, we show how teacher-student gender match affects students' aspirations. As described in the previous section, we generate two sets of estimates: students' aspirations to jobs traditionally associated with women, and students' aspirations to be either a scientist or engineer. In columns 3 and 4 of Table 3, we show results for aspirations to traditionally female jobs. The coefficients

Figure 2: Low perceived ability students' current perception of the difficulty of math, by gender of student and math teacher



Panel A: Low perceived ability students

Panel B: Non-low perceived ability students

Notes: This figure plots the proportion of students with each possible response to the prompt: "how difficult do you find your current mathematics course to be?" The plots give proportions separately by the gender of the student and of the teacher. Panel A shows this for low perceived ability students. Panel B shows this for non-low perceived ability students. Graphically, Figure A.1 gives the below-median analogue to Figure 2 and displays a similar pattern to what we describe here.

Table 3: Effects on beliefs and aspirations

	Current math class perceived as very difficult		*Aspires to jobs in art, art, design, or acting*	
	(1)	(2)	(3)	(4)
Girl x female teacher x low perceived ability	—	-0.226***	—	-0.109**
		(0.062)		(0.055)
Female teacher x low perceived ability	—	0.099**	—	-0.030
		(0.049)		(0.033)
Girl x female teacher	-0.090***	-0.038**	-0.020	-0.001
	(0.022)	(0.018)	(0.018)	(0.019)
Girl x low perceived ability	—	0.036	—	0.089***
		(0.042)		(0.033)
Girl	0.081***	0.038***	0.206***	0.188***
	(0.018)	(0.014)	(0.016)	(0.017)
Female teacher	0.009	-0.008	0.004	0.003
	(0.021)	(0.016)	(0.017)	(0.017)
Low perceived ability	—	0.532***	—	0.031
		(0.038)		(0.024)
Overall effect on LPA girls	—	-0.128***	—	-0.139***
p-value: overall effect for LPA girls = 0	—	[0.005]	—	[0.006]
Reject LPA girl effect = LPA boy effect?	—	Yes	—	Yes
Mean for non-LPA boys	0.122		0.104	
Number of observations	8,493		8,426	

Notes: This table shows results from estimating equation 1 using the dependent variables listed in the column headings. Both dependent variables are coded as (0 = No , 1 = Yes). Robust standard errors clustered at the school level are shown in parentheses. *p < 0.1, **p < 0.05, ***p < 0.01. Variation in the number of observations here and in subsequent tables stems from missing values in the dependent variable. Results are robust to restricting the sample to only observations with no missing dependent variables. Here and later, the acronym “LPA” stands for “low perceived ability.”

show that, for low perceived ability girls, being assigned a female math teacher is associated with an 11 percentage point decrease in aspiring to these jobs. The effects of being assigned a female teacher on all other groups (low perceived ability boys, all other boys, and all other girls) are at least an order of magnitude smaller and insignificant. In Table A.6 the coefficient estimate using the alternative specification of low perceived ability has the predicted sign but is not statistically significant, and in Table A.7 we see no effect on aspirations for the below-median girls assigned to female math teachers. We find no effects on girls' aspirations to be either a scientist or engineer, either for low perceived ability children or the group as a whole, and so do not present the results in tabular form. One potential explanation for this is that because of the higher selectivity of these jobs, low perceived ability children in our study are not on the margin of aspiring to work in those fields. Our framework also suggests that it may be much harder for teachers to change the beliefs of non-low perceived ability children who might be closer to this other margin, as their priors about their own ability may be firmer.

5.1.2 Math test scores

Next, we study how perceived ability impacts teacher-student gender match effects on math test scores. We present our main results in Table 4. This table follows Table 3, showing estimates first from the conventional specification, and then from the fully specified model. In column 1, we estimate a positive but not statistically significant effect of teacher-student gender match on *all* girls: $\beta_3 = 0.093$ SD ($\sigma = 0.063$). This point estimate is well within the range of estimates generated in previous work (e.g., Dee, 2007; Muralidharan and Sundararaman, 2011; Lim and Meer, 2017).

In column 2, we show the estimates from the fully specified model. These show that the test score benefits of teacher-student gender match accrue entirely to low perceived ability girls. Being assigned a female math teacher increases the math test scores of low perceived ability girls by approximately 0.45 SD.[14] The effect for the remaining (i.e., non-LPA) girls is a precise zero: the

[14] While this estimate is larger than most prior estimates of the effects of teacher-student gender match, it is for a subgroup that is particularly likely to benefit. Other studies evaluating interventions in developing countries that target

Table 4: Effects on math test scores

	(1)	(2)
Girl x female teacher	—	0.446***
x low perceived ability		(0.166)
Female teacher	—	-0.147
x low perceived ability		(0.129)
Girl x female teacher	0.093	0.007
	(0.063)	(0.054)
Girl x low perceived ability	—	-0.019
		(0.125)
Girl	0.068	0.125***
	(0.057)	(0.049)
Female teacher	0.155**	0.185***
	(0.074)	(0.068)
Low perceived ability	—	-0.806***
		(0.084)
Overall effect on LPA girls	—	0.299***
p-value: overall effect for LPA girls = 0	—	[0.008]
Reject LPA girl effect = LPA boy effect?	—	Yes
Mean for non-LPA boys	7.024	7.024
Number of observations	8,345	8,294

Notes: This table shows results from estimating equation 1 using the wave 1 midterm math test score as the dependent variable. We standardize the score variable for comparability with other relevant studies. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. *p < 0.1, **p < 0.05, ***p < 0.01.

coefficient on [girl x female teacher], i.e., β_3, is less than 0.01 SD and insignificant. The coefficient estimate for γ_2 suggests that low perceived ability boys' test scores may decline, though it is not significant and a third of the magnitude of the estimate for low perceived ability girls. In column 4 of Tables A.6 and A.6, we estimate a smaller but still positive and significant effect of teacher-student gender match on math test scores for the low perceived ability girls as classified using our our alternative definitions for low perceived ability in math.

5.1.3 Second wave results in math

We next estimate our full specification for teacher-student gender match effects in math, one year later, using the second wave of the CEPS. This data was collected one year after the first wave of the CEPS, and only from the subset of children who were in the seventh grade during the first wave. The data include perceived difficulty of current math class, job aspirations, and score on the standardized eighth grade midterm math test. We continue to estimate the impact of being assigned a female math teacher in grade 7 on these outcomes, presenting results in Table 5.

The magnitude of the estimated effects on low perceived ability girls' aspirations and test scores are of similar magnitude to estimated effects on these outcomes in the first wave, and the second wave estimate for test scores continues to be highly significant. The estimated impact on their perceived difficulty of math, however, diminishes in magnitude and ceases to be significant.

5.2 Comparing results for Math, Chinese, and English

In this section, we present our estimates of teacher-student gender match effects in Chinese and English. We present the math results from the previous section alongside, both for comparison and to highlight our framework's two key predictions. The first prediction is that the largest effects should accrue for low perceived ability students in each subject. The second is that the magnitude of the gender match effects for low perceived ability students should be smaller in English and

low performers or that work with children in particularly deprived regions report effect sizes of similar magnitudes (Banerjee and Duflo, 2007; Burde and Linden, 2013; Muralidharan et al., 2019).

Table 5: Persistence of effects of math teacher-student gender match after one year

	(1) Perceived difficulty of current math class	(2) Aspires to jobs in art and design	(3) Eighth grade midterm math test score
Girl x female teacher x low perceived ability	-0.055 (0.108)	-0.121* (0.070)	0.421*** (0.173)
Female teacher x low perceived ability	-0.022 (0.085)	0.037 (0.039)	-0.228 (0.148)
Girl x female teacher	-0.000 (0.023)	-0.003 (0.027)	0.027 (0.075)
Girl x low perceived ability	-0.057 (0.090)	0.109* (0.059)	0.025 (0.129)
Girl	0.007 (0.019)	0.234*** (0.024)	0.181*** (0.062)
Female teacher	-0.029 (0.019)	0.028* (0.015)	0.230*** (0.081)
Low perceived ability	0.364*** (0.073)	-0.001 (0.032)	-0.803*** (0.110)
Overall effect on LPA girls	-0.078	-0.084	0.193
p-value: overall effect for LPA girls = 0	[0.133]	[0.165]	[0.128]
Reject LPA girl effect = LPA boy effect?	No	Yes	Yes
Mean for non-LPA boys	0.100	0.090	7.007
Number of observations	5,205	5,209	5,282

Note: This table shows estimated impacts of teacher-student gender match in the seventh grade on outcomes measured in the eighth grade. Dependent variables are given in the column headings. The variables in columns 1 and 2 are coded as (0 = No , 1 = Yes). The test score results in column 3 are again presented in standardized (SD) units. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. *p < 0.1, **p < 0.05, ***p < 0.01.

Chinese than in math because of the additional influence of societal beliefs about differential math ability by gender.

We focus on results for two outcomes: students' beliefs about their own ability in the subject and their performance on midterm tests. We present results both for wave 1 and for wave 2. For each subject, we show results from two separate classifications of low perceived ability. The first is the core definition of low perceived ability - defining as low perceived ability those who thought the subject in question was "very hard" prior to entering middle school. The second is the alternative definition, which also includes those who thought the subject in question was "somewhat hard" prior to entering middle school. Presenting these results side-by-side shows how robust each result is to choice of classification.

We have relatively few male English teachers and relatively few girls who perceive themselves to be of low ability in Chinese. Specifically, in Chinese, 93 girls perceive themselves to be of low ability in wave 1 and 73 in wave 2, as compared with several hundred in both English and math. Given that only some of these children will have a female teacher, inference is challenging, though this constraint relaxes when we use the alternative definition of low perceived ability. In English, there are relatively few male teachers: 19 out of 215 in wave 1 and 10 out of 137 in wave 2, as opposed to the 30 to 80 male teachers in Chinese and math, depending on the sample. This limitation, as discussed in Antecol et al. (2015), restricts our ability to draw inference from the experience of having a male vs. female English teacher using these data. We show the number of male subject teachers and the number of students of each gender who perceive themselves to be of low ability in the subject (separately for each classification of low perceived ability) as additional rows at the bottom of Tables 6 and 7.

We first estimate the effects of teacher-student gender match on children's perceived difficulty of a given subject, and how this varies across subjects. We present estimates for wave 1 in Panel A of Table 6. As predicted, we observe the largest effects for math. Furthermore, the math estimates are robust across choice of classification method. The coefficients for both Chinese and English are uniformly insignificant, smaller in magnitude than for math, and diminish substantially when

Table 6: Teacher-student gender match effects on beliefs for math, Chinese, and English

Panel A: Wave 1

	Math		*Chinese*		*English*	
	(1)	(2)	(3)	(4)	(5)	(6)
	Original	Alternative	Original	Alternative	Original	Alternative
	LPA measure	LPA measure	LPA measure	LPA measure	LPA measure	LPA measure
Girl x female teacher	-0.226***	-0.136***	-0.193	-0.040	-0.164	-0.052
x low perceived ability	(0.062)	(0.042)	(0.130)	(0.038)	(0.105)	(0.041)
Female teacher	0.099**	0.038	0.121	0.027	0.129***	0.092**
x low perceived ability	(0.049)	(0.034)	(0.075)	(0.033)	(0.054)	(0.043)
Girl x female teacher	-0.038**	-0.022	-0.010	-0.013	-0.033	-0.035
	(0.018)	(0.017)	(0.013)	(0.011)	(0.034)	(0.029)
Girl x low perceived ability	0.036	0.080***	0.047	0.002	0.072	-0.080**
	(0.042)	(0.030)	(0.118)	(0.033)	(0.093)	(0.037)
Girl	0.038***	-0.001	-0.028***	-0.019*	-0.041	-0.027
	(0.014)	(0.015)	(0.012)	(0.010)	(0.033)	(0.029)
Female teacher	-0.008	-0.002	-0.001	-0.008	0.042	0.024
	(0.016)	(0.018)	(0.018)	(0.018)	(0.041)	(0.043)
Low perceived ability	0.532***	0.248***	0.465***	0.169***	0.448***	0.264***
	(0.038)	(0.024)	(0.074)	(0.027)	(0.051)	(0.041)
Overall effect on LPA girls	-0.128***	-0.098***	-0.072	-0.013	-0.035	0.039
p-value: overall effect for LPA girls = 0	[0.005]	[0.001]	[0.551]	[0.765]	[0.658]	[0.324]
Reject LPA girl effect = LPA boy effect?	Yes	Yes	No	No	No	No
Number of LPA girls	520	1799	93	596	385	1197
Number of LPA boys	442	1330	249	1167	1120	2218
Number of male teachers (out of 215)	82		44		19	
Number of observations	8493		8529		8157	

Notes: The regression specification used is given in equation 1, and the outcome variable is the 0/1 variable for whether the student finds their wave 1 math, Chinese, or English class (as indicated in the column heading) to be "very hard." Robust standard errors clustered at the school level are shown in parentheses. *p < 0.1, **p < 0.05, ***p < 0.01.

Panel B: Wave 2

	Math		*Chinese*		*English*	
	(1)	(2)	(3)	(4)	(5)	(6)
	Original	Alternative	Original	Alternative	Original	Alternative
	LPA measure	LPA measure	LPA measure	LPA measure	LPA measure	LPA measure
Girl x female teacher	-0.055	-0.045	-0.154	-0.002	-0.057	-0.030
x low perceived ability	(0.108)	(0.041)	(0.120)	(0.039)	(0.153)	(0.058)
Female teacher	-0.022	0.007	-0.024	-0.009	0.044	0.004
x low perceived ability	(0.085)	(0.039)	(0.099)	(0.035)	(0.048)	(0.045)
Girl x female teacher	-0.000	0.007	0.010	0.003	-0.064	-0.068
	(0.023)	(0.025)	(0.012)	(0.009)	(0.049)	(0.044)
Girl x low perceived ability	-0.057	-0.007	0.044	-0.061*	-0.037	-0.069
	(0.090)	(0.031)	(0.104)	(0.034)	(0.143)	(0.053)
Girl	0.007	-0.011	-0.039***	-0.022***	-0.061	-0.036
	(0.019)	(0.023)	(0.010)	(0.008)	(0.047)	(0.042)
Female teacher	-0.029	-0.028*	0.008	0.009	-0.023	-0.021
	(0.019)	(0.017)	(0.020)	(0.020)	(0.042)	(0.039)
Low perceived ability	0.364***	0.185***	0.193**	0.086***	0.286***	0.261***
	(0.073)	(0.030)	(0.089)	(0.031)	(0.039)	(0.039)
Overall effect on LPA girls	-0.078	-0.038	-0.178**	-0.011	-0.013	-0.026
p-value: overall effect for LPA girls = 0	[0.133]	[0.207]	[0.045]	[0.553]	[0.925]	[0.654]
Reject LPA girl effect = LPA boy effect?	No	No	No	No	No	No
Number of LPA girls	332	1216	73	443	222	750
Number of LPA boys	304	931	160	806	712	1470
Number of male teachers (out of 137)	47		30		10	
Number of observations	5205		5258		5042	

Notes: The regression specification used is given in equation 1, and the outcome variable is the 0/1 variable for whether the student finds their wave 2 math, Chinese, or English class (as indicated in the column heading) to be "very hard." This is estimated only on the subgroup of students for whom we have wave 2 data. Robust standard errors clustered at the school level are shown in parentheses. *p < 0.1, **p < 0.05, ***p < 0.01.

Table 7: Teacher-student gender match effects on test scores for math, Chinese, and English

Panel A: Wave 1

	Math		*Chinese*		*English*	
	(1)	(2)	(3)	(4)	(5)	(6)
	Original	Alternative	Original	Alternative	Original	Alternative
	LPA measure	LPA measure	LPA measure	LPA measure	LPA measure	LPA measure
Girl x female teacher	0.446***	0.274***	0.556*	0.197	0.298	-0.069
x low perceived ability	(0.164)	(0.108)	(0.319)	(0.128)	(0.199)	(0.142)
Female teacher	-0.147	-0.110	-0.434***	-0.196**	-0.237	-0.070
x low perceived ability	(0.128)	(0.086)	(0.171)	(0.092)	(0.175)	(0.148)
Girl x female teacher	0.007	-0.039	-0.022	-0.044	-0.042	0.040
	(0.054)	(0.060)	(0.057)	(0.065)	(0.101)	(0.093)
Girl x low perceived ability	-0.019	-0.065	-0.411	-0.094	-0.145	0.226*
	(0.124)	(0.076)	(0.280)	(0.119)	(0.195)	(0.130)
Girl	0.125***	0.212***	0.604***	0.593***	0.475***	0.328***
	(0.049)	(0.053)	(0.044)	(0.052)	(0.094)	(0.082)
Female teacher	0.185***	0.190***	0.211	0.250	0.110	0.091
	(0.068)	(0.067)	(0.154)	(0.162)	(0.139)	(0.137)
Low perceived ability	-0.806***	-0.647***	-0.151	-0.161*	-0.473***	-0.667***
	(0.084)	(0.064)	(0.140)	(0.085)	(0.165)	(0.139)
Overall effect on LPA girls	0.299***	0.165***	0.122	0.002	0.061	-0.139
p-value: overall effect for LPA girls = 0	[0.008]	[0.003]	[0.627]	[0.986]	[0.772]	[0.295]
Reject LPA girl effect = LPA boy effect?	Yes	Yes	Yes	No	No	No
Number of LPA girls	520	1799	93	596	385	1197
Number of LPA boys	442	1330	249	1167	1120	2218
Number of male teachers (out of 215)	82		44		19	
Number of observations	8294		8356		7977	

Notes: The regression specification used is given in equation 1, and the outcome variable is the student's test score on their wave 1 math, Chinese, or English class (as indicated in the column heading) midterm exam. Robust standard errors clustered at the school level are shown in parentheses. *p < 0.1, **p < 0.05, ***p < 0.01.

Panel B: Wave 2

	Math		*Chinese*		*English*	
	(1)	(2)	(3)	(4)	(5)	(6)
	Original	Alternative	Original	Alternative	Original	Alternative
	LPA measure	LPA measure	LPA measure	LPA measure	LPA measure	LPA measure
Girl x female teacher	0.421***	0.237**	0.889**	0.121	0.401	0.191
x low perceived ability	(0.173)	(0.110)	(0.385)	(0.133)	(0.431)	(0.221)
Female teacher	-0.228	-0.065	-0.348*	-0.100	-0.035	-0.149
x low perceived ability	(0.148)	(0.094)	(0.212)	(0.093)	(0.134)	(0.148)
Girl x female teacher	0.027	-0.025	-0.060	-0.054	0.047	0.008
	(0.075)	(0.080)	(0.057)	(0.058)	(0.106)	(0.129)
Girl x low perceived ability	0.025	-0.008	-0.664*	-0.024	-0.380	-0.070
	(0.129)	(0.076)	(0.349)	(0.107)	(0.430)	(0.221)
Girl	0.181***	0.273***	0.665***	0.628***	0.436***	0.402***
	(0.062)	(0.069)	(0.050)	(0.050)	(0.099)	(0.122)
Female teacher	0.230***	0.215***	0.056	0.076	0.182	0.260*
	(0.081)	(0.082)	(0.161)	(0.165)	(0.128)	(0.144)
Low perceived ability	-0.803***	-0.735***	-0.106	-0.223***	-0.590***	-0.543***
	(0.110)	(0.065)	(0.156)	(0.075)	(0.136)	(0.143)
Overall effect on LPA girls	0.193	0.172**	0.542*	0.021	0.366	0.042
p-value: overall effect for LPA girls = 0	[0.128]	[0.011]	[0.087]	[0.847]	[0.403]	[0.849]
Reject LPA girl effect = LPA boy effect?	Yes	Yes	Yes	No	No	No
Number of LPA girls	332	1216	73	443	222	750
Number of LPA boys	304	931	160	806	712	1470
Number of male teachers (out of 137)	47		30		10	
Number of observations	5282		5331		5107	

Notes: The regression specification used is given in equation 1, and the outcome variable is the student's test score on their wave 2 math, Chinese, or English class (as indicated in the column heading) midterm exam. This is estimated only on the subgroup of students for whom we have wave 2 data. Robust standard errors clustered at the school level are shown in parentheses. *p < 0.1, **p < 0.05, ***p < 0.01.

we use the broader method of classifying low perceived ability. In Panel B of Table 6 we show the results for wave 2. As shown earlier in Table 5, the measured effects on perceived difficulty of math diminish. We find no measurable effects in Chinese or English.

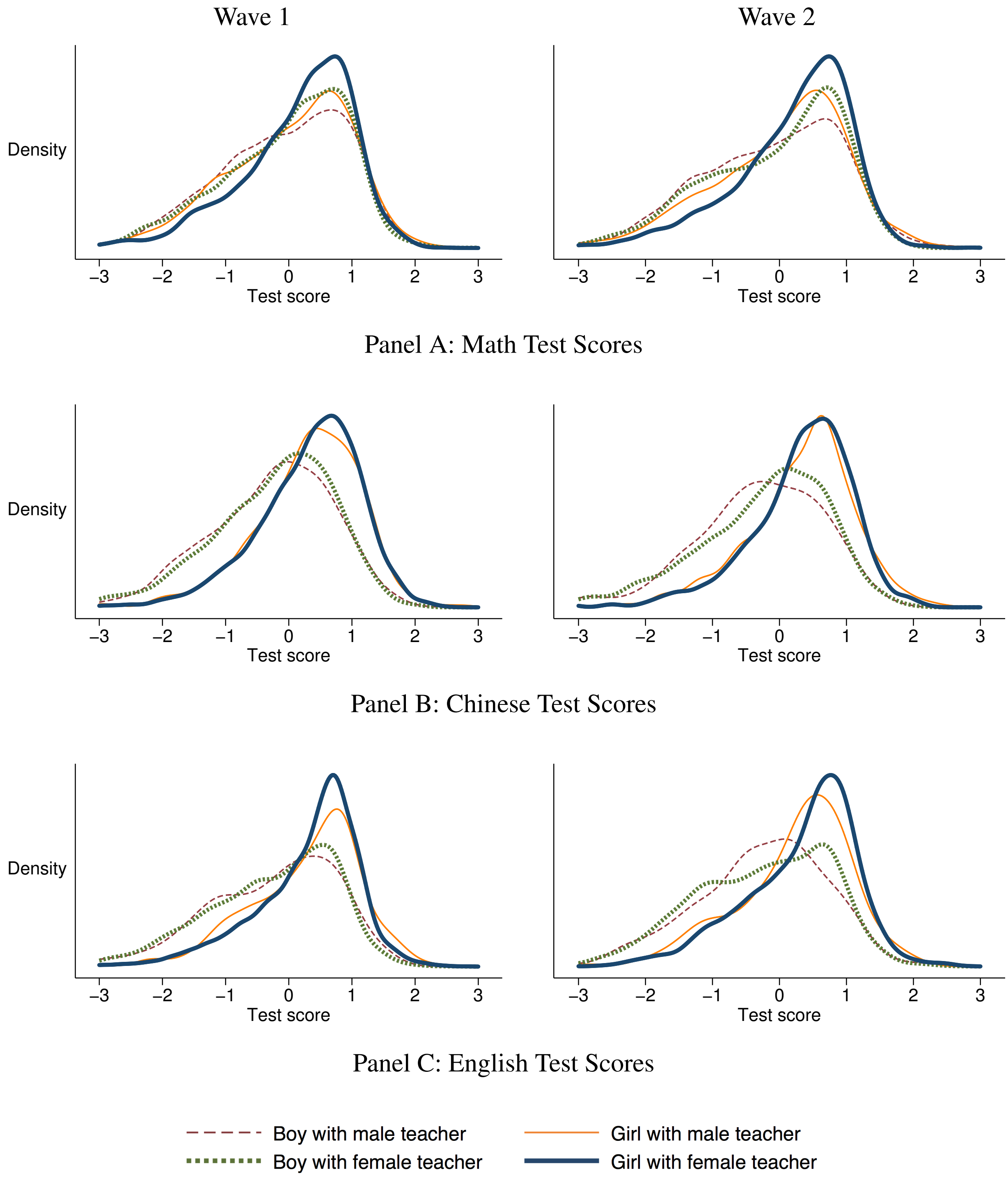
Next, we examine impacts on test scores across the three subjects and two waves. We show results for wave 1 in Panel A of Table 7. As anticipated, the effects of teacher-student gender match on low perceived ability girls' math scores are larger in magnitude and more robust than those for English or Chinese. While the coefficient for the original low perceived ability classification in Chinese (shown in column 3) is large and significant, it is estimated off of only 93 girls out of the 5,282 in the sample. Using the alternative specification of low perceived ability, the estimates given in column 4 show that the coefficient diminishes by almost an order of magnitude. For this specification, the overall effect on low perceived ability girls is a precise zero. Results for wave 2, shown in Panel B of Table 7, exhibit a similar pattern: significant effects for low perceived ability girls in math, robust across specification choice; no results for English; and the same pattern for Chinese as seen in wave 1.

In Figure 3, we show the distribution of test scores in each subject, for each possible teacher-student gender pairing, for wave 1 and wave 2 test scores. These reflect the patterns shown in both panels of Table 7: the largest effects of teacher-student gender match for girls appear in math, in the left half of the distribution. In Chinese and English, there is less or no observable difference between the scores of girls assigned to female and male teachers.

6 Mechanisms

In this section, we study the possible mechanisms driving the results we observe in the previous section. First, we conduct a series of analyses to test for the possibility that teacher behavior drives these results. Second, we test for the possibility that our findings are the result of statistical artifact, either from mean reversion or underlying differences between the determinants of boys' and girls' perceived ability. Finally, we offer our assessment of the most probable mechanisms behind the

Figure 3: Distribution of Math, Chinese, and English test scores by teacher-student gender pairing



Panel A: Math Test Scores

Panel B: Chinese Test Scores

Panel C: English Test Scores

Notes: This figure shows the distribution of students' midterm test scores in each subject in both waves by the four possible teacher-student gender pairings. We show math in the first panel (i.e., row), Chinese in the second, and English in the third. We show wave 1 test scores in the first column and wave 2 in the second. A gaussian kernel was used to generate these plots. Test scores are standardized within each grade.

empirical patterns we find.

6.1 Teacher behavior

In this section, we investigate the following possibilities: one, that female teachers give more attention to low perceived ability girls than do male teachers; two, that female teachers are merely better teachers and it is these skill differentials which drive the observed effects; three, that our effect estimates are driven by female teachers exerting more effort than male teachers; and four, that our findings are driven by differences in teaching methods between female and male teachers. Because our main empirical finding is that teacher-student gender match effects are largest and most consistently significant in math, we focus on investigating the relationship between math teacher behavior or other attributes and student outcomes.

First we investigate whether female math teachers in our sample favor girls with more praise and attention, as seen in Hoffmann and Oreopoulos (2009) and Lim and Meer (2017). The CEPS collects students' recall of how frequently their current math teacher asks them questions and of how frequently the teacher praises them in the classroom. In Table 8 we present results from estimating equation 1 using these two measures as outcome variables.[15] Our results show that while female math teachers are slightly more likely to ask students questions than male teachers, there is no evidence that they favor low perceived ability girls either with more opportunities to respond to questions or more praise.

The second possibility is that female math teachers are simply better teachers, and it is teaching skill that drives the gains we observe for girls with low perceived ability. To test for this, we generate two sets of results. First, we replace the teacher-student gender match variable (i.e., girl x female math teacher) in our estimating equation with an interaction term for girl x math teacher who won an award. We show these results in Table 9. These results do not show any evidence of "better" math teachers having a beneficial effect on low perceived ability girls' perceived difficulty

[15] Responses are coded on a four-point scale, ranging from one for "strongly disagree" to four for "strongly agree." We break this into a binary variable, mapping strongly agree and somewhat agree to one, and somewhat disagree and strongly disagree to zero. The results we show are not sensitive to recoding the middle values in either direction.

Table 8: Robustness checks - teacher attention

	(1) Is called on frequently in math class	(2) Is praised frequently in math class
Girl x female teacher x low perceived ability	0.019 (0.079)	-0.083 (0.067)
Female teacher x low perceived ability	-0.075 (0.052)	-0.009 (0.043)
Girl x female teacher	0.010 (0.025)	0.029 (0.023)
Girl x low perceived ability	0.027 (0.054)	0.038 (0.045)
Girl	-0.029 (0.022)	-0.050*** (0.018)
Female teacher	0.065* (0.034)	0.032 (0.035)
Low perceived ability	-0.101*** (0.040)	-0.143*** (0.027)
Overall effect on LPA girls	-0.057	-0.092*
p-value: overall effect for LPA girls = 0	[0.304]	[0.100]
Reject LPA girl effect = LPA boy effect?	No	No
Mean for non-LPA boys	0.635	0.513
Number of observations	8,450	8,459

Notes: This table estimates whether teacher-student gender match affects the amount of opportunities children have to participate in class (column 1) and the amount of praise they receive in class (column 2). The regression specification used here is given in equation 1. For Column 1, the dependent variable is the response, on a four point scale from one, strongly disagree, to four, strongly agree, to the prompt “the teacher calls on me frequently.” We code this as 0/1 for disagree/agree. Column 2’s dependent variable, with the same scale and coding, is the response to the prompt “the teacher often praises me.” Robust standard errors clustered at the school level are shown in parentheses. *p < 0.1, **p < 0.05, ***p < 0.01.

Table 9: Teacher aptitude

	(1) Perceived difficulty of current math class	(2) Aspires to jobs in art and design	(3) Midterm math test score
Girl x award-winning teacher	0.056	0.119*	-0.072
x low perceived ability	(0.068)	(0.065)	(0.168)
Award-winning teacher	-0.045	-0.024	0.157
x low perceived ability	(0.060)	(0.037)	(0.134)
Girl x award-winning teacher	-0.017	-0.073***	0.049
	(0.016)	(0.019)	(0.052)
Girl	-0.095*	-0.014	0.229*
x low perceived ability	(0.053)	(0.041)	(0.133)
Girl	0.022**	0.222***	0.108***
	(0.011)	(0.015)	(0.037)
Award-winning teacher	0.011	0.062***	-0.081
	(0.020)	(0.018)	(0.092)
Low perceived ability	0.604***	0.025	-0.956***
	(0.044)	(0.022)	(0.099)
Overall effect on LPA girls	0.011	0.095*	0.085
p-value: overall effect for LPA girls = 0	[0.816]	[0.067]	[0.470]
Reject LPA girl effect = LPA boy effect?	No	Yes	No
Mean for non-LPA boys	0.122	0.104	7.024
Number of observations	8,493	8,426	8,294

Notes: This table estimates the impact of math teacher aptitude (as measured by receipt of a teaching award) on student outcomes, allowing for heterogeneity by student gender and perceived ability as in earlier tables. The dependent variable is given in the column heading. Dependent variables in columns 1 and 2 are coded as (0 = No , 1 = Yes). The test score results in column 3 are presented in SD units. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. *p < 0.1, **p < 0.05, ***p < 0.01.

of math or performance in math.

We probe for this possibility further using two separate tests. First, we conduct a horse race regression, reverting to the original specification in equation 1 and adding a term to the right hand side interacting the award-winning math teacher and the teacher-student gender match dummies. We find the interaction term is negative and insignificant, while γ_3 is of similar magnitude (e.g., for the test score results, 0.3 SD or larger) and retains its statistical significance.[16] Second, we generate (but do not show) similar results for two separate specifications. One swaps the receipt of a teaching award with years of experience; the other swaps it with holding a degree from a teacher training university. These analyses also show no impact of math teacher accolades affecting either low perceived ability girl outcomes or any evidence that their inclusion in the regression changes the magnitude of our estimate of γ_3.

Next, we investigate the possibility that math teacher effort drives our results. The CEPS collects self reported time use data from teachers. We use the following data points: first, how many hours teachers spend preparing for class and grading homework, respectively. We use these as proxies for how much "effort" the teacher chooses to expend. Second, how many hours the teacher spends lecturing. We use this as a scale variable - schools determine how many classes the teacher is responsible for, which is the denominator by which we scale our raw measure of effort. We generate three measures of effort: one, [hours in preparation: hours in class]; two, [hours grading: hours in class]; and three, [(hours in preparation + hours grading): hours in class]. We use these to estimate the effect of differential effort levels, between teachers within a given school, on student outcomes. We estimate these effects in the same way we estimate the effects of teacher-student gender match, only now our independent variable of interest is the interaction of teacher effort and student gender. Because of the large volume of results this generates, we describe the results here and present the results, in tabular form, in the appendix (Tables A.8, A.9, and A.10). Our results show that, for the same dependent variables - perceived difficulty of math and math midterm exam score - we are unable to find a significant relationship between any of our

[16] For brevity, results in tabular form are available from the authors but not included in this manuscript.

teacher effort measures and child outcomes among low perceived ability girls.[17]

We also examine the correlation between teachers' use of different teaching methods and student performance. This tests for the possibility that the effects we observe are driven merely by female teachers employing different methods - e.g., engaging with students in a different way - which may affect low perceived ability girls and boys differentially. The CEPS records teachers' response to the following question: "how often do you use [teaching method]: never, sometimes, often, or always?" The question is asked separately for each of three methods, "lecturing," "small group discussion," and "interactive discussion between teacher and students." The latter two options involve more interaction between the student and teacher. We expected, a priori, for these methods to have a larger effect on the outcomes of low performing girls if teaching method does in fact drive the results in Section 5. As with the student engagement variables, there are four possible responses for how often teachers use these methods - never, sometimes, often, and always. We code these as a binary variable, with "often" and "always" mapping to one and the other responses to zero. Table A.11 shows estimates of the effect of teachers' use of these methods on perceived difficulty of math and midterm math test scores. We see no positive effect of using either method on low perceived ability girls' outcomes.

6.2 Mean reversion and the determinants of low perceived ability

In this subsection, we address the possibility that either mean reversion or some secular underlying difference between boys and girls may drive our results.

First, our main results are inconsistent with what mean reversion would predict. The central concern with mean reversion in this context is that perhaps our low perceived ability students merely had a bad draw in their sixth grade test scores and this caused them to revise their beliefs about their ability downwards. Mean reversion predicts they would be likely to have a normal draw in seventh grade (Chay et al., 2005). This would suggest that all low perceived ability students

[17] We cannot entirely exclude an alternative explanation for this pattern: teachers who expend less effort in terms of out of class hours may merely be more productive with their time.

should have a secular gain in test scores. Our estimates of a positive γ_3 in math and the difference in sign between γ_2 and γ_3 in most subjects are inconsistent with this explanation.

If the determinants of perceived ability differ between boys and girls in a way that may predict their test scores, it would influence our interpretation of γ_3. To examine this possibility, we can regress test scores on the vector of student-level predetermined characteristics and, using these coefficients, generate a predicted test score for each student. Here again, for ease of exposition and given that our main results are significant only for mathematics, we focus on the determinants of perceived ability in that subject. In Figure A.2, we plot these predicted test scores separately for boys and girls in each of the two perceived ability groups. These plots show no evidence of differences in the distribution of predicted math test scores between genders in either group.

6.3 Probable mechanisms

In this section, we describe what we think are the most probable mechanisms driving the main results of the paper. In Section 2, we outlined a conceptual framework studying the information sent from teachers to students, i.e., the "passive" effects of teacher-student gender match, as in Dee (2005). This comprises two main channels: role models and expectations.

For role models, the main mechanism we imagine is as we described in that section. Each teacher provides some information about the student's likelihood of success in the subject they teach. If the teacher shares an identity with the student, e.g., gender or ethnicity, the message sent is more credible. The intuition behind this claim is straightforward: someone who more closely resembles me is more likely to have information that pertains to me, my experience, and my potential future. Many papers have probed the intuitive underpinnings behind this. In addition to the model in the appendix of this paper, the model laid out in Gershenson et al. (2018) also formalizes this channel, explaining that via the role model channel, shared-identity teachers "lead students to update inaccurate beliefs about the returns to human capital investment." Several empirical papers have shown direct evidence of role model effects either via teachers (Bettinger and Long, 2005; Hoffmann and Oreopoulos, 2009) or other shared-identity individuals who can speak to the child's success in a given line of work. For this latter channel, Kofoed et al. (2017) and Porter and Serra (2019) study interventions in which students interact with shared-identity mentors or speakers from careers perceived as uncommon for people of that identity. These speakers tell students about their own success in these careers, providing an identity-specific example of success. In addition to providing new information, this may also directly counter societal beliefs about identity-specific levels of ability. Both papers show that interacting with these mentors or speakers can increase students' selection into and persistence in the mentor/speaker's occupation. Given that we see little evidence of differential teacher effort, teaching methods, or aptitude between male and female math teachers in our context, we argue that the role model channel is the most likely driver of the effects we observe.

The other main channel which may drive effects in our context is that of teacher expectations or beliefs. Gershenson et al. (2016) show that higher expectations for shared-identity students may contribute to teacher-student identity match effects among racial minorities in the US. A series of recent papers have shown that teacher gender bias also contributes to gender gaps in multiple settings (Alan et al., 2018; Lavy and Sand, 2018; Carlana, 2019). In our setting, differential expectations or beliefs by gender are unobservable to us, we cannot fully exclude the possibility that they contribute to the effect we observe. However, they would have to meet three conditions: first, these differential expectations would have to matter only for children who perceived themselves to be of lower ability; two, they would have to be most impactful in mathematics; and three, they would have to be conveyed in such a way that it brought up the beliefs and performance of low perceived ability girls and depressed the beliefs and performance of low perceived ability boys. Given the unobservable nature of these expectations, we cannot directly test for this possibility in our context. Nonetheless, the match between our conceptual framework and results, alongside other recent evidence on the existence and nature of role model effects, suggests that the role models channel is, at the very least, one important contributor to the patterns we observe here.

7 Conclusion

Teacher-student identity match is an important policy tool for educators interested in raising the academic performance of children. In our paper, we show conceptually how positive teacher-student identity match effects are most likely to accrue among students who are uncertain about their own ability, and why they should be greater for students whose beliefs about their ability have been artificially suppressed by exposure to societal messages about ability by identity. Using nationally representative data from Chinese middle schools, we show empirical results consistent with these predictions. We find that low perceived ability girls benefit from being assigned a female math teacher, that low perceived ability boys are harmed by being assigned a female math teacher, and no evidence of effects for other children. These patterns also appear, but are less prominent, in English and Chinese, subjects without the same societal belief about boys' superiority. Together, these results suggests that the intersection of a child's beliefs about themself and societal beliefs about ability by gender has important implications for predicting the incidence and size of teacher-student gender match effects.

Our findings have important implications for both policy and research. For education policy, one problem that educators have historically faced is how to assign the often limited number of teachers from certain identities to children who share these identities. Our results suggest prioritizing the assignment of these teachers to those students who both share their identity and perceive themselves to be of low ability in the subject being taught. For research, our results suggest study of how student beliefs relate to teacher-student identity match effects for other groups that face similar societal beliefs about ability, such as racial and ethnic minorities in the US and beyond. Our findings also highlight the possibility that student beliefs may serve as an important indicator of receptivity to information-based educational interventions. More broadly, a key message of our paper is that the informational environment a child faces - including, but not limited to, societal beliefs - plays an important role in many children's educational careers and trajectories.

REFERENCES

Alan, Sule, Seda Ertac, and Ipek Mumcu, "Gender stereotypes in the classroom and effects on achievement," *Review of Economics and Statistics*, 2018, *100* (5), 876–890.

Ambuehl, Sandro and Shengwu Li, "Belief updating and the demand for information," *Games and Economic Behavior*, 2018, *109*, 21–39.

Antecol, Heather, Ozkan Eren, and Serkan Ozbeklik, "The effect of teacher gender on student achievement in primary school," *Journal of Labor Economics*, 2015, *33* (1), 63–89.

Banerjee, Abhijit V. and Esther Duflo, "The economic lives of the poor," *Journal of Economic Perspectives*, 2007, *21* (1), 141.

Bettinger, Eric P and Bridget Terry Long, "Do faculty serve as role models? The impact of instructor gender on female students," *American Economic Review*, 2005, *95* (2), 152–157.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian, "Gender stereotypes about intellectual ability emerge early and influence children's interests," *Science*, 2017, *355* (6323), 389–391.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Beliefs about gender," *American Economic Review*, 2019, *109* (3), 739–73.

Bruhn, Miriam and David McKenzie, "In pursuit of balance: Randomization in practice in development field experiments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 200–232.

Burde, Dana and Leigh L Linden, "Bringing education to Afghan girls: A randomized controlled trial of village-based schools," *American Economic Journal: Applied Economics*, 2013, *5* (3), 27–40.

Camerer, Colin, "Individual decision making," *Handbook of Experimental Economics*, 1995.

Carlana, Michela, “Implicit stereotypes: Evidence from teachers’ gender bias,” *The Quarterly Journal of Economics*, 2019, *134* (3), 1163–1224.

Carman, Katherine Grace and Lei Zhang, “Classroom peer effects and academic achievement: Evidence from a Chinese middle school,” *China Economic Review*, 2012, *23* (2), 223–237.

Carrell, Scott E and James E West, “Does professor quality matter? Evidence from random assignment of students to professors,” *Journal of Political Economy*, 2010, *118* (3), 409–432.

_ , Marianne E Page, and James E West, “Sex and science: How professor gender perpetuates the gender gap,” *Quarterly Journal of Economics*, 2010, *125* (3), 1101–1144.

Chay, Kenneth Y, Patrick J McEwan, and Miguel Urquiola, “The central role of noise in evaluating interventions that use test scores to rank schools,” *American Economic Review*, 2005, *95* (4), 1237–1258.

Cheryan, Sapna, “Understanding the paradox in math-related fields: Why do some gender gaps remain while others do not?,” *Sex Roles*, 2012, *66* (3-4), 184–190.

Cho, Insook, “The effect of teacher–student gender matching: Evidence from OECD countries,” *Economics of Education Review*, 2012, *31* (3), 54–67.

Dee, Thomas S, “Teachers, race, and student achievement in a randomized experiment,” *Review of Economics and Statistics*, 2004, *86* (1), 195–210.

_ , “A teacher like me: Does race, ethnicity, or gender matter?,” *American Economic Review*, 2005, *95* (2), 158–165.

_ , “Teachers and the gender gaps in student achievement,” *Journal of Human Resources*, 2007, *42* (3), 528–554.

Eble, Alex and Feng Hu, “How important are beliefs about gender differences in math ability? Transmission across generations and impacts on child outcomes,” *CDEP-CGEG Working Paper*, 2019, *53.*

Ellison, Glenn and Ashley Swanson, “The gender gap in secondary school mathematics at high achievement levels: Evidence from the American Mathematics Competitions,” *Journal of Economic Perspectives*, 2010, *24* (2), 109–128.

Gärdenfors, Peter, *Belief Revision*, Vol. 29, Cambridge University Press, 2003.

Gershenson, Seth, Cassandra Hart, Joshua Hyman, Constance Lindsay, and Nicholas W Papageorge, “The long-run impacts of same-race teachers,” *National Bureau of Economic Research WP 25254*, 2018.

_ , Stephen B Holt, and Nicholas W Papageorge, “Who believes in me? The effect of student–teacher demographic match on teacher expectations,” *Economics of Education Review*, 2016, *52*, 209–224.

Gong, Jie, Yi Lu, and Hong Song, “The effect of teacher gender on students’ academic and noncognitive outcomes,” *Journal of Labor Economics*, 2018, *36*, 743–778.

Graham, Patricia Albjerg, “Black teachers: A drastically scarce resource,” *The Phi Delta Kappan*, 1987, *68* (8), 598–605.

Hansen, Ben B and Jake Bowers, “Covariate balance in simple, stratified and clustered comparative studies,” *Statistical Science*, 2008, pp. 219–236.

He, Leshui, Stephen L Ross et al., “Classroom peer effects and teachers: Evidence from quasi-random assignment in a Chinese middle school,” *Human Capital and Economic Opportunity Global Working Group Working Paper 2017-014*, 2017.

Hoffmann, Florian and Philip Oreopoulos, “A professor like me the influence of instructor gender on college achievement,” *Journal of Human Resources*, 2009, *44* (2), 479–494.

Hu, Feng, “Do girl peers improve your academic performance?,” *Economics Letters*, 2015, *137*, 54–58.

Jayachandran, Seema, "The roots of gender inequality in developing countries," *Annual Review of Economics*, 2015, *7* (1), 63–88.

Kipnis, Andrew B, *Governing educational desire: Culture, politics, and schooling in China*, University of Chicago Press, 2011.

Kofoed, Michael S et al., "The effect of same-gender and same-race role models on occupation choice: Evidence from randomly assigned mentors at West Point," *Journal of Human Resources*, 2017, pp. 0416–7838r1.

Lavy, Victor and Analia Schlosser, "Mechanisms and impacts of gender peer effects at school," *American Economic Journal: Applied Economics*, 2011, *3* (2), 1–33.

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

_ **and Rigissa Megalokonomou**, "Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study," *NBER Working Paper 26021*, 2019.

Lim, Jaegeum and Jonathan Meer, "The impact of teacher-student gender matches: Random assignment evidence from South Korea," *Journal of Human Resources*, 2017, *52* (4), 979–997.

_ **and** _ , "Persistent Effects of Teacher-Student Gender Matches," *Journal of Human Resources*, Forthcoming.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian, "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 2019, *109* (4), 1426–60.

_ **and Ketki Sheth**, "Bridging education gender gaps in developing countries: The role of female teachers," *Journal of Human Resources*, 2016, *51* (2), 269–297.

_ **and Venkatesh Sundararaman**, "Teacher performance pay: Experimental evidence from India," *Journal of Political Economy*, 2011, *119* (1), 39–77.

Niederle, Muriel and Lise Vesterlund, "Explaining the gender gap in math test scores: The role of competition," *Journal of Economic Perspectives*, 2010, *24* (2), 129–144.

Nollenberger, Natalia, Núria Rodríguez-Planas, and Almudena Sevilla, "The math gender gap: The role of culture," *American Economic Review*, 2016, *106* (5), 257–61.

Porter, Catherine and Danila Serra, "Gender Differences in the Choice of Major: The Importance of Female Role Models," *Working Paper*, 2019.

Rodríguez-Planas, Núria and Natalia Nollenberger, "Let the girls learn! It is not only about math...it's about gender social norms," *Economics of Education Review*, 2018, *62*, 230 – 253.

Shih, Margaret, Todd L Pittinsky, and Nalini Ambady, "Stereotype susceptibility: Identity salience and shifts in quantitative performance," *Psychological Science*, 1999, *10* (1), 80–83.

Spencer, Steven J, Claude M Steele, and Diane M Quinn, "Stereotype threat and women's math performance," *Journal of Experimental Social Psychology*, 1999, *35* (1), 4–28.

Tsui, Ming, "Gender and mathematics achievement in China and the United States," *Gender Issues*, 2007, *24* (3), 1–11.

Wilson, W.J., *The truly disadvantaged: The inner city, the underclass, and public policy*, University of Chicago Press, 2012.

Xu, Di and Qiujie Li, "Gender achievement gaps among Chinese middle school students and the role of teachers' gender," *Economics of Education Review*, 2018, *67*, 82–93.

Appendix - for online publication only

Appendix A: Appendix tables

Table A.1: Summary statistics for teachers

	Math		*Chinese*		*English*	
	(1)	(2)	(3)	(4)	(5)	(6)
	Female	Male	Female	Male	Female	Male
Age	36.7	39.5	36.0	40.5	35.8	38.5
Attended a normal university	0.92	0.98	0.96	0.98	0.88	0.84
Years of teaching experience	15.7	18.6	15.0	19.1	14.7	17.5
Holds a senior rank in the school	0.24	0.22	0.20	0.23	0.14	0.21
Won award						
At the province or national level	0.14	0.13	0.14	0.18	0.09	0.16
At the city level	0.41	0.48	0.40	0.57	0.30	0.37
Highest educational level						
Associate college or below	0.08	0.20	0.05	0.32	0.09	0.32
Part-time four year university	0.33	0.37	0.39	0.39	0.40	0.32
Full-time four year university	0.54	0.41	0.48	0.27	0.46	0.37
Master's degree or higher	0.05	0.02	0.07	0.02	0.04	0.00
Number of teachers	132	82	168	44	191	19

Notes: This table compares observable teacher characteristics across teacher gender in each subject. This table also uses only data from the main estimation sample in the paper, described in Section 4.1. The differences in total number of teachers across subjects reflect missing data: we are missing teacher gender in one case for math, three cases for Chinese, and five cases for English.

Table A.2: Background characteristics, summarized by gender and perceived ability

	Perceived ability			
	Low		*Not low*	
	(1)	(2)	(3)	(4)
	Girls	Boys	Girls	Boys
Age	13.50	13.52	13.13	13.26
Ethnic minority	0.23	0.19	0.11	0.10
Holds agricultural hukou	0.56	0.64	0.47	0.49
Number of siblings	1.06	0.93	0.72	0.63
Low household income / poor	0.30	0.30	0.16	0.19
Father's years of schooling	9.47	9.30	10.86	10.68
Mother's years of schooling	8.41	8.41	10.21	9.92
Number of observations	536	471	3,934	4,351

Panel A: Math

	Perceived ability			
	Low		*Not low*	
	(1)	(2)	(3)	(4)
	Girls	Boys	Girls	Boys
Age	13.26	13.47	13.17	13.27
Ethnic minority	0.21	0.15	0.12	0.11
Holds agricultural hukou	0.39	0.60	0.49	0.50
Number of siblings	0.99	0.78	0.75	0.65
Low household income / poor	0.21	0.31	0.17	0.19
Father's years of schooling	9.49	9.62	10.72	10.59
Mother's years of schooling	8.02	8.89	10.04	9.82
Number of observations	94	260	4,369	4,553

Panel B: Chinese

	Perceived ability			
	Low		*Not low*	
	(1)	(2)	(3)	(4)
	Girls	Boys	Girls	Boys
Age	13.58	13.58	13.11	13.17
Ethnic minority	0.20	0.18	0.11	0.08
Holds agricultural hukou	0.63	0.64	0.46	0.46
Number of siblings	1.03	0.88	0.72	0.58
Low household income / poor	0.32	0.28	0.16	0.17
Father's years of schooling	9.21	9.38	10.89	10.98
Mother's years of schooling	7.93	8.29	10.27	10.34
Number of observations	395	1,177	4,007	3,582

Panel C: English

Notes: This table shows group-specific means for the low perceived ability girls and boys in our sample and, separately, for those who are not low perceived ability. We show results for each subject in a separate panel.

Table A.3: One-by-one tests for orthogonality of teacher gender and predetermined characteristic

	Math	Chinese	English
	(1)	(2)	(3)
Number of siblings	-0.013	-0.008	0.005
	(0.011)	(0.018)	(0.018)
Household is poor	-0.011	0.043	-0.011
	(0.021)	(0.037)	(0.028)
Female	-0.010	0.041*	-0.019
	(0.007)	(0.023)	(0.017)
Age	0.032	-0.007	0.014
	(0.025)	(0.018)	(0.017)
Ethnic minority	0.018	-0.060	0.061
	(0.016)	(0.112)	(0.098)
Holds agricultural hukou	0.026	0.008	0.002
	(0.038)	(0.015)	(0.013)
Low perceived ability in Math	0.030	-0.092	0.058
	(0.022)	(0.061)	(0.061)
Low perceived ability in English	-0.016	-0.060	0.047
	(0.041)	(0.045)	(0.032)
Low perceived ability in Chinese	0.040	-0.093	0.095
	(0.032)	(0.064)	(0.063)
Mother's education level			
Middle school	0.022	-0.007	0.037**
	(0.046)	(0.033)	(0.018)
High/technical school	-0.005	0.043*	-0.053***
	(0.035)	(0.025)	(0.022)
College or above	-0.023	-0.005	-0.016
	(0.062)	(0.034)	(0.018)
Father's education level			
Middle school	0.007	-0.021	0.027
	(0.015)	(0.030)	(0.028)
High/technical school	-0.010*	-0.001	-0.011
	(0.006)	(0.024)	(0.019)
College or above	-0.014	0.018	-0.045***
	(0.061)	(0.035)	(0.016)

Notes: This table shows coefficient and standard error estimates from regressing teacher gender on the predetermined teachers characteristics listed in the first column, one-by-one (e.g., each coefficient is from a separate regression), controlling also for grade-by-school fixed effects. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A.4: Tests for gender-specific teacher quality

	(1) Math	(2) Chinese	(3) English
Age	-0.018	-0.023	-0.009
	0.030	0.016	0.028
Has B.A.	0.055	0.131	-0.028
	0.249	0.146	0.198
Went to teachers' college	-0.222	0.055	-0.003
	0.216	0.425	0.175
Years of experience	0.015	0.013	0.005
	0.027	0.014	0.027
Won award at province level	0.161	0.304	-0.222
	0.387	0.221	0.179
Won award at city level	-0.108	-0.096	0.072
	0.255	0.141	0.126
Number of observations	207	207	202
R-squared	0.70	0.84	0.78
Joint test F-statistic	0.25	0.75	0.32
[p-value]	[0.96]	[0.61]	[0.93]

Notes: This table shows coefficient and standard error estimates from regressing teacher gender on the predetermined teachers characteristics listed in the first column (along with grade-by-school fixed effects as in our main estimating equation) and conducting a Wald Test for their joint significance, similar to the results shown in Table 2 for student characteristics. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A.5: Test for randomization, low perceived ability students only

	(1) Math	(2) Chinese	(3) English
Number of siblings	0.001	-0.014	-0.004
	(0.013)	(0.034)	(0.007)
Household is poor	0.014	0.005	-0.004
	(0.026)	(0.042)	(0.015)
Female	-0.015	0.003	-0.000
	(0.016)	(0.027)	(0.014)
Age	-0.006	-0.000	0.012
	(0.007)	(0.042)	(0.010)
Ethnic minority	0.026	-0.077	0.004
	(0.023)	(0.078)	(0.008)
Holds agricultural hukou	-0.042	-0.027	0.000
	(0.032)	(0.042)	(0.010)
Mother's education level			
Middle school	-0.008	-0.022	0.001
	(0.022)	(0.040)	(0.006)
High/technical school	0.043	-0.097	0.009
	(0.038)	(0.070)	(0.006)
College or above	-0.069	-0.048	0.023**
	(0.065)	(0.060)	(0.012)
Father's education level			
Middle school	-0.012	0.097***	0.011
	(0.027)	(0.040)	(0.009)
High/technical school	-0.041	0.147*	0.009*
	(0.045)	(0.081)	(0.006)
College or above	0.149***	0.092*	-0.013
	(0.061)	(0.056)	(0.016)
Number of observations	850	308	1,309
R-squared	0.85	0.88	0.74
Joint test F-statistic	1.55	0.76	0.87
[p-value]	[0.12]	[0.69]	[0.58]

Notes: This table shows results from three separate omnibus regressions of teacher gender (=1 if female in the subject listed in the column heading) on the set of independent variables listed in the first column along with grade-by-school fixed effects as in our main estimating equation. This follows Bruhn and McKenzie (2009) and Antecol et al. (2015). We show the coefficients for the variable with the robust standard errors below. Here, we restrict the sample to only the low perceived ability students in each subject.

Table A.6: Effect of teacher-student gender match on beliefs and performance using alternative perceived ability definition

	(1) Perceived difficulty of current math class	(2) Aspires to jobs in art and design	(3) Midterm math test score
Girl x female teacher x low perc- eived ability (alternate definition)	-0.136*** (0.042)	-0.059* (0.035)	0.274*** (0.108)
Female teacher x LPA alternate definition	0.038 (0.034)	0.012 (0.018)	-0.110 (0.086)
Girl x female teacher	-0.022 (0.017)	0.009 (0.021)	-0.039 (0.060)
Girl x LPA alternate definition	0.080*** (0.030)	0.075*** (0.025)	-0.065 (0.076)
Girl	-0.001 (0.015)	0.168*** (0.018)	0.212*** (0.053)
Female teacher	-0.002 (0.018)	-0.001 (0.017)	0.190*** (0.067)
Low perceived ability (alternate definition)	0.248*** (0.024)	0.009 (0.013)	-0.647*** (0.064)
Overall effect on LPA girls	-0.098***	-0.048*	0.165***
p-value: overall effect for LPA girls = 0	[0.001]	[0.083]	[0.003]
Reject LPA girl effect = LPA boy effect?	Yes	Yes	Yes
Mean for non-LPA boys	0.094	0.100	7.137
Number of observations	8,493	8,426	8,294

Notes: This table replicates the analyses of Tables 3 and 4, only using the alternative definition of whether the student perceives themself to be of low ability in math. The dependent variable in question is given in the column heading. Dependent variables in columns 1 and 2 are coded as (0 = No , 1 = Yes). The test score results in column 3 are presented in SD units. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A.7: Effect of teacher-student gender match on beliefs and performance using below median test score instead of perceived ability

	(1) Perceived difficulty of current math class	(2) Aspires to jobs in art and design	(3) Midterm math test score
Girl x female teacher	-0.092***	0.006	0.111*
x below median	(0.029)	(0.034)	(0.057)
Female teacher	-0.006	-0.006	0.087***
x below median	(0.026)	(0.019)	(0.033)
Girl x female teacher	-0.041**	-0.014	0.012
	(0.018)	(0.022)	(0.029)
Girl x below median	0.061***	0.024	0.056
	(0.022)	(0.026)	(0.046)
Girl	0.052***	0.186***	0.047**
	(0.015)	(0.019)	(0.024)
Female teacher	0.018	0.004	0.061**
	(0.021)	(0.018)	(0.030)
Below median (b-m)	0.197***	0.045***	-1.653***
	(0.019)	(0.016)	(0.029)
Overall effect on b-m girls	-0.098***	0.000	0.198***
p-value: overall effect for LPA girls = 0	[0.002]	[1.000]	[0.000]
Reject b-m girl effect = b-m boy effect?	Yes	No	Yes
Mean for above median boys	0.069	0.085	7.757
Number of observations	8,300	8,251	8,345

Notes: This table replicates the analyses of Tables 3 and 4, only splitting the sample on whether the student was below the median in their math teacher-student gender pairing instead of whether the student perceives themself to be of low ability in math. The dependent variable in question is given in the column heading. Dependent variables in columns 1 and 2 are coded as (0 = No , 1 = Yes). The test score results in column 3 are presented in SD units. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A.8: Relationship between math teacher effort and student outcomes: effort measure 1

	(1)	(2)	(3)
	Perceived	Enrolled in	Midterm
	difficulty	after-school	math
	of current	math	test
	math class	tutoring	score
Hours prep: hours in class x low perceived ability x girl	0.059	0.029	-0.062
	(0.056)	(0.031)	(0.123)
Hours prep: hours in class x low perceived ability	-0.034	-0.013	0.152**
	(0.033)	(0.023)	(0.076)
Hours prep: hours in class x girl	-0.011	-0.013	0.051
	(0.008)	(0.011)	(0.032)
Girl x low perceived ability	-0.136*	-0.043	0.283
	(0.077)	(0.051)	(0.177)
Girl	0.027**	0.056***	0.075
	(0.013)	(0.017)	(0.048)
Hours prep: hours in class	0.024***	-0.033***	-0.097*
	(0.009)	(0.010)	(0.058)
Low perceived ability	0.620***	0.012	-1.055***
	(0.050)	(0.028)	(0.108)
Overall effect on LPA girls	0.025	0.016	0.090
p-value: overall effect for LPA girls = 0	[0.558]	[0.555]	[0.478]
Reject LPA girl effect = LPA boy effect?	No	No	No
Mean for non-LPA boys	0.122	0.210	7.024
Number of observations	8,429	8,410	8,230

Notes: This is a robustness test to see how teacher effort, instead of teacher gender, affects student outcomes, allowing for heterogeneity by student gender and perceived ability as in earlier tables. Dependent variables are given in the column headings. The dependent variables in columns 1 and 2 are coded as (0 = No , 1 = Yes). The test score results in column 3 are presented in SD units. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A.9: Relationship between math teacher effort and student outcomes: effort measure 2

	(1) Perceived difficulty of current math class	(2) Enrolled in after-school math tutoring	(3) Midterm math test score
Hours grading: hours in class x low perceived ability x girl	0.016	0.020	-0.099
	(0.044)	(0.048)	(0.130)
Hours grading: hours in class x low perceived ability	-0.059	-0.004	0.149*
	(0.037)	(0.029)	(0.087)
Hours grading: hours in class x girl	-0.015*	0.004	0.051*
	(0.008)	(0.012)	(0.030)
Girl x low perceived ability	-0.092	-0.034	0.319*
	(0.069)	(0.053)	(0.186)
Girl	0.032***	0.039**	0.075*
	(0.013)	(0.017)	(0.045)
Hours grading: hours in class	0.003	-0.005	0.075
	(0.013)	(0.015)	(0.057)
Low perceived ability	0.646***	0.002	-1.050***
	(0.055)	(0.033)	(0.122)
Overall effect on LPA girls	-0.043*	0.016	0.050
p-value: overall effect for LPA girls = 0	[0.096]	[0.608]	[0.593]
Reject LPA girl effect = LPA boy effect?	No	No	No
Mean for non-LPA boys	0.122	0.210	7.024
Number of observations	8,429	8,410	8,230

Notes: This is a robustness test to see how teacher effort, instead of teacher gender, affects student outcomes, allowing for heterogeneity by student gender and perceived ability as in earlier tables. Dependent variables are given in the column headings. The dependent variables in columns 1 and 2 are coded as (0 = No , 1 = Yes). The test score results in column 3 are presented in SD units. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. *p < 0.1, **p < 0.05, ***p < 0.01.

Table A.10: Relationship between math teacher effort and student outcomes: effort measure 3

	(1) Perceived difficulty of current math class	(2) Enrolled in after-school math tutoring	(3) Midterm math test score
Hours prep + grading: hours in class x low perceived ability x girl	0.025 (0.033)	0.017 (0.021)	-0.053 (0.076)
Hours prep + grading: hours in class x low perceived ability	-0.032 (0.023)	-0.005 (0.015)	0.105** (0.051)
Hours prep + grading: hours in class x girl	-0.009* (0.005)	-0.003 (0.007)	0.033* (0.018)
Girl x low perceived ability	-0.127 (0.089)	-0.049 (0.054)	0.334 (0.212)
Girl	0.034*** (0.014)	0.048*** (0.019)	0.057 (0.050)
Hours prep + grading: hours in class	0.011* (0.006)	-0.016** (0.008)	-0.020 (0.035)
Low perceived ability	0.653*** (0.063)	0.009 (0.034)	-1.121*** (0.134)
Overall effect on LPA girls p-value: overall effect for LPA girls = 0	-0.008 [0.718]	0.012 [0.469]	0.052 [0.409]
Reject LPA girl effect = LPA boy effect?	No	No	No
Mean for non-LPA boys	0.122	0.210	7.024
Number of observations	8,429	8,410	8,230

Notes: This is a robustness test to see how teacher effort, instead of teacher gender, affects student outcomes, allowing for heterogeneity by student gender and perceived ability as in earlier tables. Dependent variables are given in the column headings. The dependent variables in columns 1 and 2 are coded as (0 = No , 1 = Yes). The test score results in column 3 are presented in SD units. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. *p < 0.1, **p < 0.05, ***p < 0.01.

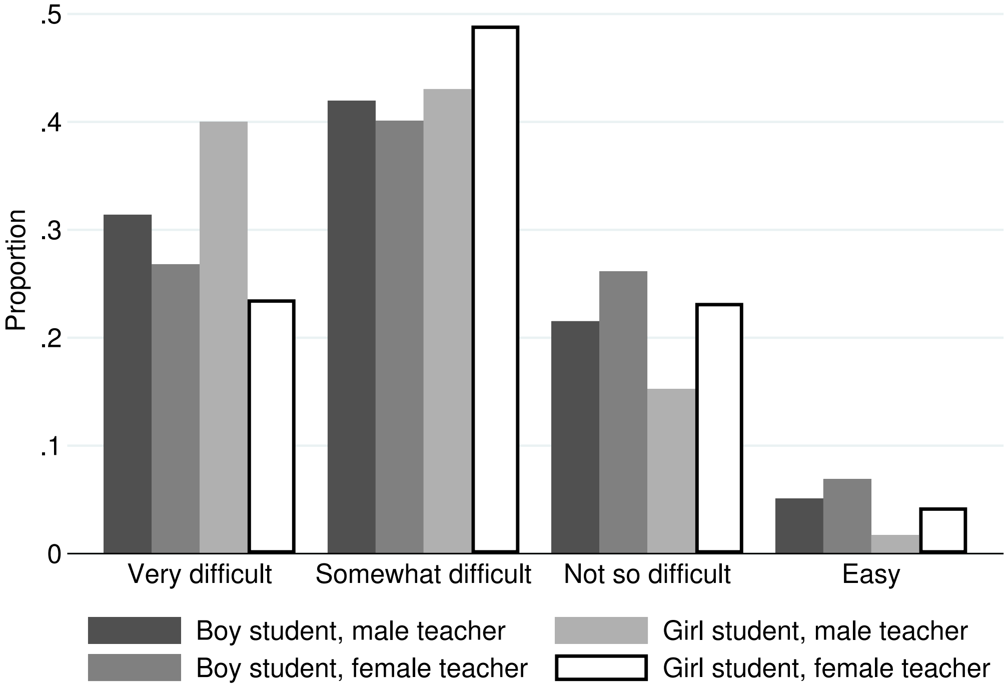
Table A.11: Relationship between math teacher teaching method and student outcomes

	Discuss in small groups		*Students and teacher "interactively" discuss*	
	(1)	(2)	(3)	(4)
	Perceived difficulty of current math class	Midterm math test score	Perceived difficulty of current math class	Midterm math test score
Uses teaching method	0.018	-0.103	0.071	-0.157
x girl x LPA	(0.065)	(0.153)	(0.069)	(0.155)
Uses teaching	0.028	0.001	-0.023	-0.115
method x LPA	(0.047)	(0.123)	(0.052)	(0.131)
Uses teaching method	-0.011	0.022	-0.030	-0.052
x girl	(0.014)	(0.049)	(0.023)	(0.062)
Girl x LPA	-0.079	0.254**	-0.117***	0.310***
	(0.052)	(0.122)	(0.045)	(0.109)
Girl	0.020*	0.119***	0.038*	0.167***
	(0.011)	(0.033)	(0.021)	(0.054)
Uses teaching method	0.017	-0.046	0.004	-0.014
	(0.021)	(0.095)	(0.025)	(0.116)
Low perceived	0.570***	-0.889***	0.596***	-0.801***
ability (LPA)	(0.028)	(0.089)	(0.036)	(0.107)
Overall effect on LPA girls	0.046	-0.102	0.048	-0.272**
p-value: overall effect for LPA girls = 0	[0.400]	[0.343]	[0.326]	[0.022]
Reject LPA girl effect = LPA boy effect?	No	No	No	No
Mean for non-LPA boys	0.122	7.024	0.122	7.024
Number of observations	8,474	8,275	8,468	8,268

Notes: This is a robustness test to see how the teacher's teaching method, instead of teacher gender, affects student outcomes, allowing for heterogeneity by student gender and perceived ability as in earlier tables. Dependent variables are given in the column headings. The dependent variables in columns 1 and 3 are coded as (0 = No , 1 = Yes). The test score results in columns 2 and 4 are presented in SD units. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in equation 1. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

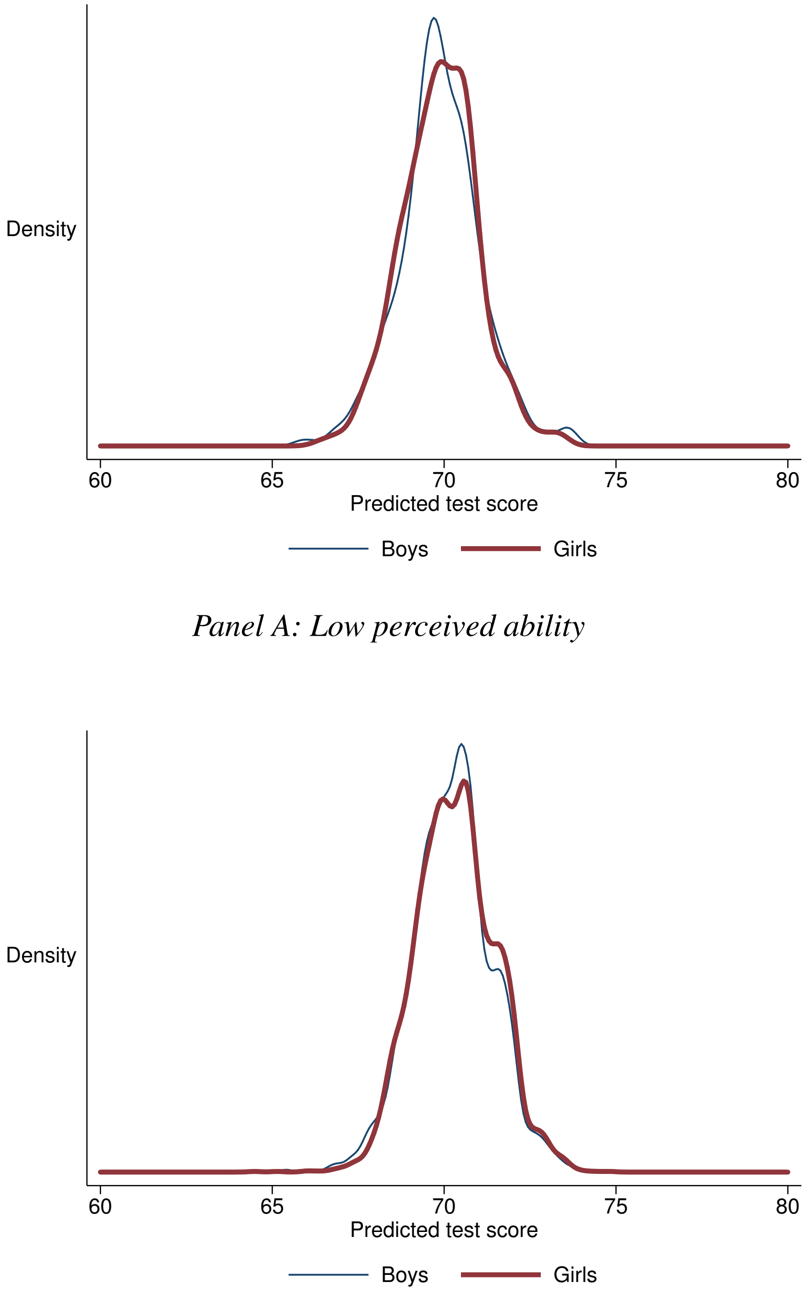
Appendix B: Appendix figures

Figure A.1: Effect of math teacher-student gender match on student beliefs, for those below within-group median test score



Notes: This figure shows the same analysis as reported in Panel A of Figure 2, only limiting the sample instead to those below the within-group median math test score.

Figure A.2: Predicted math test score distributions, by perceived ability



Panel A: Low perceived ability

Panel B: Not low perceived ability

Notes: To generate these figures, we regress math test scores on the vector of student-level predetermined characteristics and, using these coefficients, generate a predicted test score for each student. We then plot these using a gaussian kernel for each perceived ability-gender group.

Appendix C: Description of balanced assignment rule

Assume that one middle school has a total of 200 incoming seventh-grade students, who will be assigned to five classes. Students are first ranked by their total scores on primary school graduation examinations and then are assigned to classes according to their score ranks in an alternating way - for the first five students, student 1 is assigned to class 1, student 2 is assigned to class 2, and so on until student 5. Then, student 6 is assigned to class 5, student 7 to class 4, an on until student 10 is assigned to class 1. Then the original order repeats, so that student 11 is assigned to class 1, student 12 to class 2, and on until student 15. At student 16, the order once again reverses, and so on, so as to avoid bifurcation of classrooms (that is, avoiding the case where the best and worst students are placed together in some classrooms and mid-level performers are placed together in others). This is described nicely in He et al. (2017), who, along with Hu (2015) and Gong et al. (2018), also exploit this quasi-random assignment of students to classes in Chinese middle schools.