COLUMBIA | SIPA
Center for Development Economics and Policy

CDEP-CGEG WORKING PAPER SERIES

CDEP-CGEG WP No. 43

Stereotypes, Role Models, and the Formation of Beliefs

Alex Eble and Feng Hu

December 2017

COLUMBIA | SIPA
Center on Global Economic Governance

Stereotypes, Role Models, and the Formation of Beliefs

Alex Eble and Feng Hu*

December 2017

Abstract

Information affects beliefs, which in turn determine investment decisions. Because human capital exhibits dynamic complementarity, early sources of information play a crucial role in its formation. We study how information from stereotypes and role models influences children's beliefs, aspirations, investment, and academic performance. A model of investment under uncertainty predicts that role models should have the greatest effect for children facing stereotypes who are also on the margin of giving up on themselves. We exploit random assignment of students to classes in a nationally-representative dataset of Chinese middle schools to test the model's main predictions and address potential alternative explanations.

*Eble: Teachers College, Columbia University. Email: eble@tc.columbia.edu Hu: School of Economics and Management, University of Science and Technology Beijing. Email: feng3hu@gmail.com. The authors are grateful to Anjali Adukia, Joe Cummins, John Friedman, Morgan Hardy, Asim Khwaja, Ilyana Kuziemko, Bentley MacCleod, Derek Neal, Randy Reback, Jonah Rockoff, Judy Scott-Clayton, and Felipe Valencia for generous input, as well as seminar audiences at Columbia, Fordham, the 2017 IZA transatlantic meeting, and PacDev. We also acknowledge financial support from the National Natural Science Foundation of China (grant nos. 71373002, 71420107023). Key words: gender; belief formation; stereotypes; human capital; cognitive skills; behavioral economics. JEL codes: D83; I24; J16; O15.

1 Introduction

As a person goes through life, her beliefs are shaped by what she comes in contact with. These beliefs inform her investment decisions and, more broadly, her aspirations for the future. A long line of both empirical and theoretical work has explored the importance of this causal chain in the formation of human capital (Becker, 1975; Jensen, 2010; Akerlof and Kranton, 2002; Lybbert and Wydick, 2016b; Genicot and Ray, 2017). Because human capital formation is characterised by dynamic complementarity, this link from information to beliefs and on to human capital investment decisions is particularly important in the early stages of life.

A crucial implication of this relationship is that early misinformation about ability can have long-lasting negative consequences. If telling a child she is of low ability in a given subject reduces her relevant investment in period one, dynamic complementarity implies that in period two she will be relatively less able in the subject, bearing out the prediction of the misinformation and reinforcing its message. A common source of such misinformation is negative stereotypes about ability by gender and ethnicity (Steele and Aronson, 1995; Steele, 2003). Recent evidence suggests that gender stereotypes affect the interests and time use decisions of both girls and boys as early as age seven (Bian et al., 2017) and that such stereotypes may lead to underrepresentation of women and minorities in several scientific fields where such misinformation persists (Leslie et al., 2015).

In this paper, we study how children's beliefs are affected by stereotypes and role models, and how this influences aspirations, investment in skills, and academic performance. We first develop a simple stylised model based on that of Genicot and Ray (2017). In our model, children are uncertain about their own ability and the returns to schooling. Stereotypes provide erroneous group-specific signals about ability. Role models also send signals about children's ability, and if the role model shares an identity with the child, the signal to noise ratio of her message is higher. Teachers serve as role models because they have frequent contact with students and occupy a place of authority and expertise

(Latane, 1981; Bettinger and Long, 2005). Our model predicts that the greatest benefits of encountering a role model accrue to children who face a negative stereotype and who are on the margin of deciding not to invest in human capital, i.e., on the margin of "giving up."

We then take the predictions of our model to data from a nationally representative set of Chinese middle schools. In this setting, there is widespread belief among children that boys are better than girls in learning math. Due to a revision of China's compulsory education laws in 2006 which banned the tracking of entering middle school students to different classes based on academic performance, there is also random assignment of students to classrooms. This allows us to estimate the causal effects of different classroom configurations, a method one of us has used in previous work to study peer effects (Hu, 2015) and which has subsequently been used in other studies (He et al., 2017; Gong et al., Forthcoming).

We estimate the effects of teacher-student gender match on girls' and boys' beliefs, behaviors, and academic outcomes, generating subgroup-specific estimates by whether the student perceives her/himself to be of low ability in math. In line with the predictions of our model, we find low perceived ability girls assigned to female math teachers are 20 percentage points less likely to perceive math 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 are 9 percentage points more likely to enroll in mathematics tutoring (baseline 15%). These girls also perform 0.45 standard deviations better on a standardised math test. While this effect estimate is larger than commonly seen in the literature, it is for a subgroup; our overall estimate of the effect of teacher-student gender match on girls' test scores is is 0.09 SD, well within the range of estimates generated in prior work (cf. Dee, 2007; Muralidharan and Sundararaman, 2011; Lim and Meer, Forthcoming). Other work studying how to raise performance among low performers or in particularly needy areas in less developed countries has found similarly large effects (Banerjee et al., 2007; Burde and Linden, 2013). Consistent with an ancillary prediction of our model and the psychological concept of "identity threat" (Steele et al., 2002; Sherman et al., 2013), we find negative effects of being assigned a female math teacher on low perceived ability boys' perceived difficulty and investment. We see no gender-specific benefit for assigning a same-gendered teacher to boys or girls who do not perceive themselves to be of low ability in math.

We then conduct a series of analyses to disentangle the different mechanisms which may drive the effects we observe. Our results suggest that the main mechanism at work is female math teachers serving as role models for low perceived ability girls, protecting against the harmful effects of negative gender stereotypes. We provide three sets of analyses to support this claim. First, as described above, our model's empirical predictions are borne out in our data - we see large effects of teacher-student gender match on the beliefs, aspirations, and investment behavior of low perceived ability girls, some evidence of negative effects for low perceived ability boys who are assigned a female math teacher, and no effects for other students. Second, we conduct a battery of tests for the possibility that variation in teaching methods, aptitude, or effort between male and female teachers drive the effects we observe, and find no evidence that any of these differentially affect low perceived ability students. Third, we test for the possibility that female teachers give extra attention, either praise or opportunity to speak in class, to low perceived ability girls, and find no evidence of such differential treatment.

Our work builds on two active lines of inquiry in economics. The first is the budding set of studies on the formation of aspirations and beliefs and their role in forward-looking decisions, particularly those related to human capital formation. Several recent papers, both empirical (Bernard et al., 2014; Lybbert and Wydick, 2016a; Ross, 2016; Kofoed et al., 2017) and theoretical (Akerlof and Kranton, 2000, 2002; Bénabou and Tirole, 2011; Lybbert and Wydick, 2016b), have studied the role of aspirations in affecting investment behavior. To this work, we add empirical evidence of an important informational channel through which beliefs and aspirations can be influenced. Our results are consistent with both the hypothesis of Wilson (2012) and a key prediction of the model in Genicot and Ray (2017); namely, that informational shocks about oneself (as opposed to about the world, as in Jensen, 2010) may induce changes in aspirations, which in turn lead to changes in investment and outcomes. Our analysis contributes novel evidence on the protective effect role models can have on the formation of beliefs and aspirations for students facing negative stereotypes at a crucial stage of child development.

The second is the set of studies estimating the effects of teacher-student identity match on the performance of stereotyped-against individuals (e.g., Bettinger and Long 2005; Dee 2007; Carrell et al. 2010; Fairlie et al. 2014; Gershenson et al. 2016; Muralidharan and Sheth 2016; Gong et al. Forthcoming). While this literature has argued for and shown evidence of several possible mechanisms driving the largely positive effects found, we provide the first direct evidence we are aware of in support of a specific mechanism, the power of role models to shape beliefs, aspirations, and investment behavior, in driving the positive effects of teacher-student gender match on student test scores. This complements recent work showing that shared-identity teachers may serve as role models, shaping career choices of stereotyped-against college students in the US (Carrell et al., 2010; Kofoed et al., 2017). Furthermore, unlike many previous studies on teacher-student gender match, we study a setting where, overall, girls perform slightly better than boys in mathematics. Nonetheless, negative gender norms about girls' math ability persist, and are strongest among the low perceived ability girls. It is precisely among these girls that we see the largest impact of being assigned a female math teacher.

The rest of this paper is structured as follows. In Section 2 we outline our conceptual framework to motivate the focus of the paper on low perceived ability girls. Section 3 describes the setting we study. Section 4.1 outlines our data sources and provides summary statistics of our main variables. Section 4.2 introduces our empirical strategy and presents results for tests of our main identifying assumptions. Section 5 presents our main empirical results estimating the effects of being assigned a female math teacher on student beliefs, aspirations, investment, and performance. Section 6 investigates possible mechanisms for these effects and discusses the limitations of our study. The final section concludes.

2 Conceptual framework and empirical motivation

In this section, we first write down a conceptual framework, drawing on Genicot and Ray (2017), that formalizes the intuition of how stereotypes and role models may affect the formation of beliefs. This generates predictions we test later in the paper. We then conduct a simple distributional difference test using our data which further motivates the empirical focus on low perceived ability girls.

Both across countries and in our Chinese data, girls express a disproportionate lack of confidence in their own ability in math as well as in the math ability of their gender (Beilock et al., 2010; OECD, 2015). The empirical literature in psychology demonstrates that these beliefs directly contribute to worse performance among women via two channels. First, anxiety because of "stereotype threat" (Shih et al., 1999; Spencer et al., 1999; Niederle and Vesterlund, 2010; Cheryan, 2012) could lead to lower performance on high stakes math assessments, which would in turn affect later life outcomes. Second, negative gender norms may exert downward pressure on beliefs in the returns to investment, causing girls to invest less effort, enthusiasm, and time in studying for math (Bian et al., 2017). This leads to lower performance and thus a confirmation of the once-erroneous content of the stereotype. A corollary of these findings is that the presence of a female teacher in the same subject as the stereotypical belief could change girls' views about the potential positive returns to their effort in math. The female teacher provides a credible (by virtue of shared gender) example of the returns to such effort (Carrell et al., 2010; Wilson, 2012; Genicot and Ray, 2017) which, in turn, may change girls' willingness to exert effort in the subject area (Nixon and Robinson, 1999; Beaman et al., 2009; Gunderson et al., 2012). Evidence from psychology also suggests that such an example could lead to an increase in students' academic motivation and expectations (Nixon and Robinson, 1999; Gershenson et al., 2016).

To formalise this intuition, we place our analysis in the context of a canonical two period model of consumption and savings. Individuals face the following consumption problem:

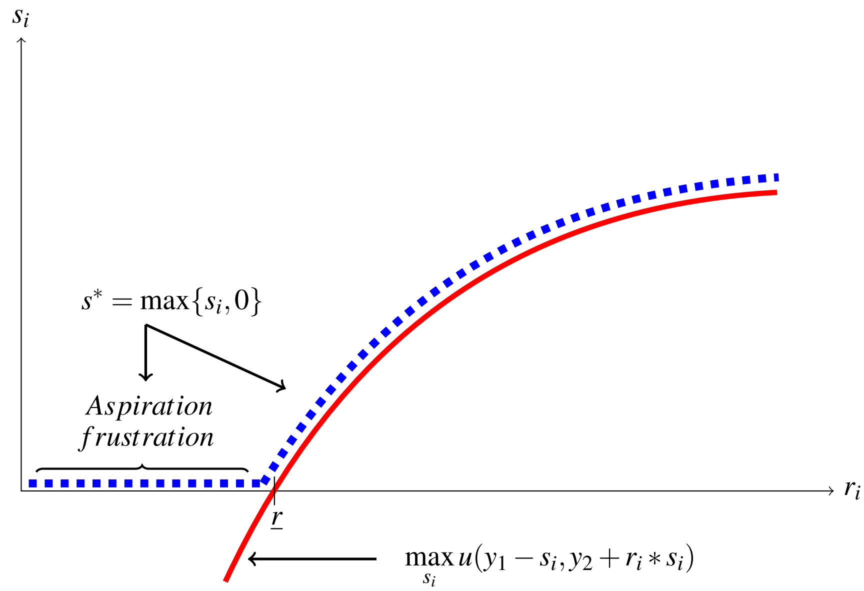
$$\max_{s_i} u(y_1 - s_i, y_2 + r_i * s_i) \tag{1}$$

We assume utility from consumption is concave in both periods. Defining terms, y_t is income in period t, s_i is the savings of individual i in period 1, and r_i is individual i's belief about her return on saving in period 1, earned in period 2. Instead of modeling savings as money, here the savings technology is investment in human capital, a combination of effort and time exerted beyond the bare minimum in class, on homework, and in seeking out extra assistance via tutoring, normalised to zero when the student exerts the minimum possible effort. We assume r_i is a function of the individual's endowment and the informational environment the individual faces, which can include information gleaned from parents, peers, the media, societal stereotypes, and so forth.

We introduce stereotypes and link to the Genicot and Ray model through r_i. We assume there is a part of the positive support of r, $[0, \underline{r})$, over which the interest rate does not justify investment. For some of the individuals whose perceived return on investment falls in this range, their low r_i is caused by incorrect information, that is, exposure to either stochastic shocks or systematic biases such as stereotypes. These individuals will rationally but sub-optimally choose not to invest, part of what Genicot and Ray call "aspiration frustration." In Figure 1, we depict this static part of our model graphically.

The effect of new information on the update to r_i will depend on two parameters of the signal, its credibility and the difference between the individual's prior and the new information provided by the signal. Informally, for girl students who perceive themselves to be of sufficiently low ability in math that they fall into the aspiration frustration part of the support of r_i, being assigned a female math teacher provides a signal of the potential for returns to investment in human capital that is credible because of the teacher's shared gender, and far from these girls' priors, which have been formed both by gender stereotypes and the low signals about ability received up to that point. For other students, the combina-

Figure 1: Visual depiction of the model



tion of positive signals (girls and boys who are doing well in math) and a lack of negative stereotypes (all boys) lead to the prediction that being assigned a same-gendered math teacher is unlikely to cause these individuals to update r_i in response to a female teacher as dramatically as will low perceived ability girls.

Formally, individual i proceeds through life gaining new information about r_i from her environment and experiences. For convenience we divide beliefs about r_i into a discrete variable A_i:

$$A_i = \begin{cases} L & if \;\; r_i < \underline{r} \\ H & if \;\; r_i \geq \underline{r} \end{cases} \tag{2}$$

Our object of interest is a set of conditional probabilities $P(H|G^i, P^i, T^i)$, where the conditions relate to the gender of the student, her/his perceived ability in mathematics, and the gender of the middle school math teacher. We define student gender as $G^i \in \{G^g, G^b\}$, where g and b indicate the student is a girl or a boy, respectively. We define student perceived ability as $P^i \in \{P^l, P^h\}$, where l and h indicate the student perceives herself to be of low or high ability, respectively. We define teacher's gender as $T^i \in \{T^f, T^m\}$, where

f and m indicate the teacher is female or male, respectively. In the data we see that $P(H|G^g) < P(H|G^b)$, that is, girls have lower perceived ability in math than boys[1]. We make three further assumptions that allow us to generate three predictions to test in our data.

Assumption 1: all teachers attempt to send the message that, with enough investment, a student can succeed in math.

Assumption 2: $P(G^x, T^y|L) < P(G^x, T^y|H)$ for $(x,y) \in \{(b,m);(g,f)\}$ that is, encountering a same-gendered math teacher delivers a signal that the individual is more likely to be H than L. We base this on the notion that teachers' role of authority gives their message some credibility and that a message sent by a teacher who shares an identity with the student provides a higher signal to noise ratio than that sent by a teacher with whom the student has no shared identity[2]. This generates a mapping from $P(H|G^g)$ to $[P(H|G^g, T^f) - P(H|G^g)]$ that has a right-skewed inverse-U shape. We show an example of this in Figure A.1.

Assumption 3: $P(H|G^g) > \tilde{H}$, where $\tilde{H}$ is some level of H strictly greater than zero, ensuring that girls do not perceive themselves to be so unlikely to be H that they will not update in response to a signal (e.g., they are not in the leftmost portion of Figure A.1).

Prediction 1: $P(H|G^g, T^f) - P(H|G^g) > P(H|G^b, T^*) - P(H|G^b)$. In prose, we predict that girls assigned to a female math teacher should update their prior on their ability to productively invest more than boys assigned to either a female or male math teacher. This is a direct consequent of girls' lower perceived ability, Assumption 3, and Bayes' rule.

Prediction 2: $P(H|G^g, P^l, T^f) - P(H|G^g, P^l) > P(H|G^g, P^h, T^f) - P(H|G^g, P^h)$, that is, we predict low-perceived ability girls will make larger updates to their prior than high perceived ability girls in response to encountering a female teacher. This prediction comes from Assumptions 2 and 3, and is derived from a basic tenet of information theory: information that is relatively new to the receiver generates a larger update to the prior than it would for receivers for whom the information is less novel[3]. Seen through the lens of Bayesian

[1] Given that girls have slightly better math test scores than boys, we argue that at least some of the gap in perceived ability is due to negative gender stereotypes.

[2] These predictions are also derived in work on the psychological concept of "Social Impact Theory" (Latane, 1981).

[3] That is, low perceived ability girls exposed to an example of success see it as more novel than do high

updating, high perceived ability girls have a much higher $P(H)$ than low perceived ability girls. As shown in Figure A.1, high perceived ability girls thus update far less than low perceived ability girls from the same information.

Corollary: depending on the proximity of r_i to $\underline{r}$, we should also see increases in s_i and academic performance among the low perceived ability girls assigned to a female math teacher.

Prediction 3: $P(H|G^b, P^l, T^f) - P(H|G^b, P^l) < 0$, that is, being assigned a female math teacher will reduce low perceived ability boys' belief in their ability to productively invest in themselves. This is derived from the psychological concept of identity threat, which refers to the negative response (low performance, reduced effort) that occurs when members of a privileged group see a threat to the status quo (Scheepers and Ellemers, 2005). In our context, the existing stereotype posits that boys are better at learning math than girls. Low perceived ability boys, confronted with the dual threats of an increase in the difficulty of math when they enter middle school and the appearance of a female math teacher, may interpret these as signals that threaten their perception of the status quo that, as boys, they are better than girls in learning math.

Next, we examine two features of our data - the distribution of perceived ability by gender and the distribution of math test scores - as preliminary tests of our assumptions and our main prediction, respectively. Despite performing no worse on math tests than boys, girls are nearly 10 percentage points more likely to report that they find math at least somewhat difficult. Furthermore, girls to the left of the median math test score are 15 percentage points more likely than girls to the right of the median to believe that boys are better at math than girls (54% vs. 39%). We offer this as prima facie evidence that negative gender stereotypes affect girls in our data, particularly low perceived ability girls.

Next, in Figure 2, we show a kernel density plot of math test scores for the four different teacher-student gender pairings ($G^g : T^f$, $G^g : T^m$, $G^b : T^f$, and $G^b : T^m$). Girls assigned a female math teacher outperform all other pairings, but only in (roughly) the left half of the

perceived ability girls, who in themselves already have an example of success.

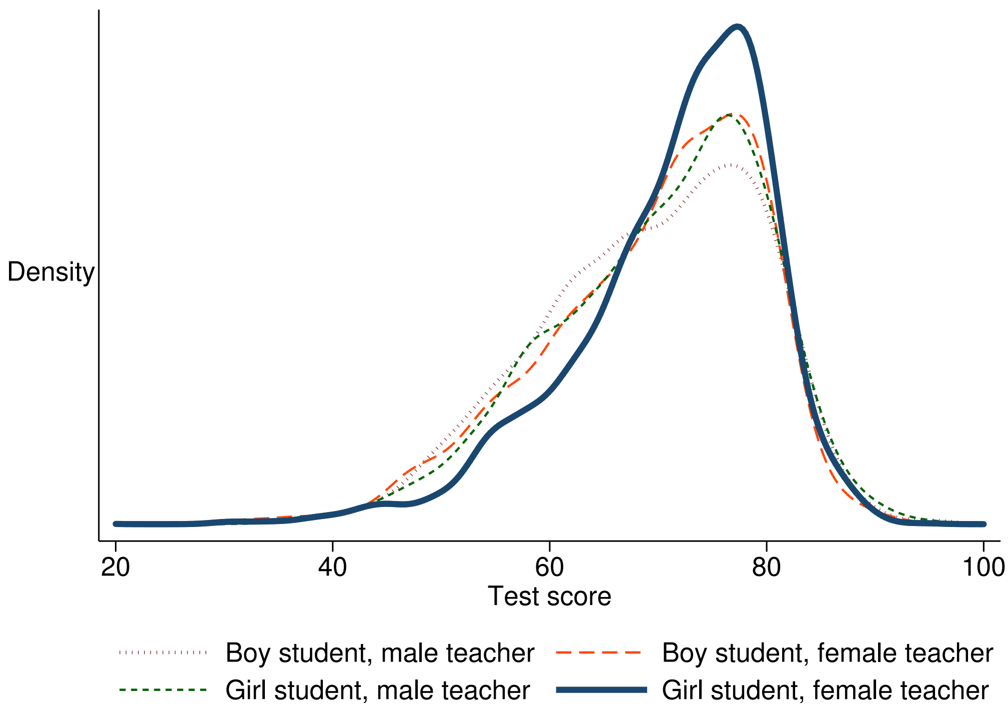
distribution. A Kolmogorov-Smirnov test rejects the equality of the $G^g : T^f$ distribution from the combined distribution of the test scores of students in other teacher-student gender pairings with a p-value of less than 0.001, and, as we will discuss further in Section 5, quantile regressions show substantial gains in the first through third deciles. As well as being in line with our main theoretical predictions, these results also suggest that we should look among students in the left tail of the ability distribution for the potential impacts of teacher-student gender match.

3 Setting

China's 1986 compulsory education law mandated that all children receive nine years of free compulsory education, including six years of primary schooling (the first to sixth grades) and three years of middle school education (seventh to ninth grade). Until the late 1990s, primary school graduates were required to attend an entrance examination to be eligible to enter middle school (Lai et al., 2011; Carman and Zhang, 2012). At the turn of the millennium, middle schools were prohibited from selecting students based on academic merit and the middle school entrance examination was later cancelled. In the same spirit, tracking of students to different classes based on demonstrated ability or academic performance has been banned in middle schools since a subsequent compulsory education law was issued in 2006.

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. Primary school graduates are assigned to a neighborhood middle school according to local educational authorities' regulations, e.g., districting. In the first system, they are then randomly assigned to classes by lottery or another quasi-

Figure 2: Distribution of math test scores by teacher-student gender pairing



Notes: This figure plots the distribution of students' scores on math midterm examinations by the four possible teacher-student gender pairings. The sample is restricted to the estimation sample as described in Section 4. A gaussian kernel was used to generate the density plots. A Kolmogorov-Smirnov test rejects equality of the distributions of test scores between two groups: girls paired with a female teacher and the combined distribution of students in all other teacher-student gender configurations. Test scores are standardised within each grade within a given school so that ten points is one standard deviation and the mean is 70.

random method[4]. 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). Hu (2015), He et al. (2017), and Gong et al. (Forthcoming) exploit this level of random assignment of students to classes and provide explanations of the nature of classroom assignment[5]. In Appendix C, we provide a description of this type of assignment rule, borrowing from He et al. (2017). 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.

In our empirical work, we will exploit these two methods of assigning students to classes as providing potentially quasi-random matching of student gender to teacher gender. We note, however, that this system is not implemented with perfect fidelity, particularly as students move beyond the first grade of middle school, i.e., from the seventh grade to the eighth and ninth. Unlike in many western countries, where admission to high school or university is either according to residence or based on multiple dimensions (e.g., grades and teacher recommendations), China’s high school admissions system relies almost exclusively on entrance examination scores (Zhang, 2016). Furthermore, the promotion of middle school administrators is largely determined by their school’s students’ performance in the high school entrance examination. More specifically, promotion is often awarded according to the annual number of graduates admitted to elite high schools. As a result, de-

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

[5]This paper was written at the same time as Gong et al. (Forthcoming), who use the same data and identification strategy, but analyse both a different subsample of the data and a different set of questions. The two papers serve as complements; Gong et al. show the salutary effects of female teachers on female students’ academic outcomes and non-cognitive skills aggregated across subjects (math, English, and Chinese). The first main point of departure is that we restrict our sample to mathematics classes, focusing on the machinery of how role models and anti-girl stereotypes in math interact and for whom. The second point of departure is that, informed by our theory, we decompose effects by whether the student is of low perceived ability. Using their sample (all classes, not just mathematics) and removing the low perceived ability interactions, we are able to reproduce their findings. To the extent that there are qualitative differences in results (i.e., in teacher attention and praise), they stem from our focus on low perceived ability students and our deliberate restriction to mathematics classes.

spite the banning of class tracking some middle schools assign students to classes based on their academic performance in order to better prepare top students for the entrance examination. Along with this sorting, school administrators may channel better teachers and more resources to classes with higher-ability students to maximise the chances that some of these students place in the best high schools. This practice is more common in the eighth and ninth grades than in the seventh and means that after their first semester or year of middle school, students may be reassigned to different classes based on their academic performance even if they are randomly assigned at the beginning of the seventh grade. In this analysis, as in Hu (2015), we restrict our attention to students randomly assigned to classes in the 7th grade and in those schools where random assignment of students to classes is maintained throughout middle school.

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 the identification strategy we use, stating and testing our identifying assumptions.

4.1 Data sources

The main data source we use in this paper is the baseline wave[6] of 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 labour market activities. The baseline survey of the CEPS adopted 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

[6] At the time of writing, this is the only available wave of the study.

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 and their families, as well as detailed information on students' beliefs, aspirations, and time use. It also collects administrative school records on students' midterm test scores in the following three compulsory subjects: math, Chinese, and English. The scores are standardised 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. (Forthcoming) argue that blinded grading is common in these particular tests. In the footnote, we make a slightly weaker argument: that low stakes math exam scores are unlikely to be substantially biased by teacher gender, and it is even less likely that they will be biased for low perceived ability girls assigned to female teachers[7].

The teacher questionnaire contains rich information on teacher characteristics, including teachers' age, gender, education levels, years of teaching experience, whether the teacher graduated from a university for teachers, whether the teacher holds a senior professional rank, and whether the teacher has won any teaching awards at various levels. The survey also contains information on the subject and the class the teacher taught during the 2013-2014 academic year. We limit most of our analyses to the matched math teacher-student dataset.

[7] First, midterm exams in mathematics offer less scope for manipulation than English or Chinese because they are graded on more objective criteria (e.g., was the number produced the correct answer?). Second, in Section 6 we present evidence (Table 5) that female teachers do not favor girls or low perceived ability girls either with more opportunities to respond to questions or with more praises in the classroom, suggesting that female math teachers may also not favor low performing girls in grading.

The survey also collects data on the assignment mechanism used to assign students to classrooms, collected both from school principals and homeroom teachers[8]. The options are 1) tracking, 2) assignment according to students' household registration location, 3) either literally random assignment ("sui ji", meaning 'by chance') or according to the average-equilibrating algorithm described above, or 4) 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. Among those schools, one third reassigned students based on past academic performance when they entered the eighth or ninth grade. In our analysis, we will treat assignment to class as random for seventh graders in those schools reporting use of either purely random assignment or the average-equalization algorithm to assign seventh-grade students to classes, and for 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. If this assumption is valid, our approach allows us to causally estimate the effect of teacher gender on student outcomes[9].

Table A.1 presents summary statistics for students by gender for those students randomly assigned to classrooms. 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 tests administered in class.

Table A.2 shows summary statistics for teachers in the classrooms studied in Table A.1. In our data, 39% of the students are taught by male math teachers, alleviating the challenge faced in Antecol et al. (2015) where there was an insufficient number of male teachers to

[8] 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.

[9] This is assumption also investigated in Hu (2015).

draw strong conclusions from some of the comparisons made. Female math teachers are on average younger and less experienced than their male counterparts. However, female teachers appear to be more qualified than their male counterparts in terms of education and proportion having won a teaching award at the province or national level[10].

The significant differences in characteristics between girls and boys and between female and male math teachers above may reflect certain gender-specific patterns at the region or school level. For instance, girls and female teachers may be more likely to come from urban schools. In the next subsection, we show evidence that our empirical approach reduces the risk of potential bias stemming from such heterogeneity between teachers between schools. Specifically, our empirical strategy compares male and female teachers within a grade within a school. Below, we show that these observed differences attenuate dramatically and cease to be significant at this level of comparison.

4.2 Empirical strategy

In this subsection we first discuss our approach to estimating the effects of being assigned a female math teacher on female and on male students. We then test the identifying assumptions we must satisfy in order to interpret our coefficient estimates causally.

In this paper we exploit the random assignment of students to classrooms to estimate the impact of teacher-student gender match on students' beliefs, aspirations, investment behavior, and performance on in-school examinations. We estimate a reduced form regression equation, controlling for grade-by-school fixed effects and a vector of observable, predetermined characteristics at the child and teacher levels. Specifically, to determine whether teacher-student gender match differentially affects the gender gap in outcomes of interest, we estimate the following equation using CEPS data:

$$\begin{aligned} Y_{icgj} = \beta_0 + \beta_1 FS_{icgj} + \beta_2 FT_{cgj} + \beta_3 (FS_{icgj} * FT_{cgj}) + \gamma_0 LPA_{icgj} + \gamma_1 (LPA_{icgj} * FS_{icgj}) + \\ \gamma_2 (LPA_{icgj} * FT_{cgj}) + \gamma_3 [LPA_{icgj} * (FS_{icgj} * FT_{cgj})] + \beta_4 SC_{icgj} + \beta_5 TC_{cgj} + \eta_{gj} + \varepsilon_{icgj} \end{aligned} \tag{3}$$

[10] 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.

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 teacher in class c in grade g of school j is female. LPA_{icgj} is an indicator equal to one if the student perceives herself to be of low ability. SC_{icgj} is a vector of predetermined characteristics at the student level, TC_{cgj} is a similar vector for teachers, η_{gj} is a grade-by-school fixed effect, and ε_{icgj} is a robust standard error, clustered at the school level[11] to allow for heteroskedasticity and arbitrary serial correlation across students within a given school. 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 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, and whether she or he has 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 assigned to either a female or male math 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 beliefs, investment, or achievement. We test this assumption later in this section.

Our model generates clear predictions for three parameters. The first is γ_3, which we interpret as a quasi-experimental estimate of the effect of assigning a female math teacher on low-perceived ability girls relative to the effect for low perceived ability boys, i.e., the effect of teacher-student gender match on the "gender gap" for low perceived ability children (Dee, 2007; Muralidharan and Sundararaman, 2011; Lim and Meer, Forthcoming). Prediction 1 is that this coefficient should be non-zero and point in the direction of reducing

[11] All of our results continue to hold if we instead cluster at the (less conservative) classroom level.

the gender gap, e.g., positive for test scores and negative for perceived difficulty of math. The second is β_3, the effect on the gender gap for non-low perceived ability children. Prediction 2 of our model is that this coefficient should be substantially smaller in magnitude than γ_3. Prediction 3 of our model pertains to γ_2. This coefficient captures the effect on all low perceived ability students of being paired with a female teacher, using those assigned to a male teacher as the comparison. By virtue of the inclusion of γ_3, γ_2 is also the entire effect of being assigned a female math teacher on low perceived ability boys. The model predicts γ_2 and γ_3 to differ in sign.

There are several parameters of ancillary interest that are derived from different combinations of the coefficients we estimate in Equation 3, and we will explicitly address a few of these in the results. First, $\gamma_2 + \gamma_3$ yields the total effect on low perceived ability girls of being assigned a female teacher relative to low perceived girls assigned a male teacher (that is, the effect of being assigned a female teacher on low ability students plus the effect of being assigned a female teacher on low ability girl students). Second, $\beta_3 + \gamma_3$ yields the total effect on the gender gap (i.e., the comparison group is now all boys, not only low perceived ability boys).

If our assumption of orthogonality is satisfied, estimating Equation 3 using OLS should recover unbiased estimates of these parameters. To test this assumption (that within a grade within a given school, the match of student gender to teacher gender is randomly assigned), we follow Antecol et al. (2015), regressing math teacher gender on the same set of observable, predetermined student and family characteristics described above that we control for in our main empirical specification[12]. We conduct two regressions - one without any fixed effects, and a second with the grade-by-school fixed effects we use in our main empirical specification. For each regression we present coefficient 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 columns 1 and 2 of Table 1. With the inclusion of grade-by-school fixed effects, our F-test fails to reject the null that the regressors are together not

[12]This method is also discussed in Hansen and Bowers (2008) and Bruhn and McKenzie (2009).

significant predictors of teacher gender (column 2). This result supports our main identifying assumption that students' observable predetermined background characteristics are balanced along the gender of math teachers within the same grade in a given school[13].

Another descriptive comparison of interest is teacher quality across genders. This paper aims to investigate the effect of female math teachers on student achievement. To ensure that we are isolating the effect of gender, 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). Table A.4 reports the estimation results for conducting a similar empirical test to that in Table 1, only conducting the analysis at the teacher level. These predetermined characteristics include age, a dummy for having earned a full-time bachelor's degree or higher qualification, a dummy for having attended a normal university, years of teaching experience, and two dummies for winning teaching awards at different levels. After conditioning on grade-by-school fixed effects, we again fail to reject the null that within a grade within a school, these characteristics are not jointly predictive of the teacher's gender. While we cannot rule out the possibility that in some cases influential parents or individuals successfully lobbied to be placed with a better 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.

As we rely on teachers' and principals' reports of whether they use tracking or random assignment, it may also be the case that some schools which report using random assignment in fact use tracking. Deliberate misreporting of tracking as "random" 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 under a tracking system wherein the administrators are seeking to maximise the performance of the best students.

The CEPS asks students how difficult they found learning math in the sixth (and final)

[13] 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.

Table 1: Test for randomization

	Full sample		*Low perceived ability*	
	(1)	(2)	(3)	(4)
Number of siblings	-0.021	-0.006	-0.026	0.001
	(0.016)	(0.006)	(0.025)	(0.013)
Household is poor	-0.053	0.005	-0.100**	0.014
	(0.033)	(0.013)	(0.046)	(0.026)
Female	0.000	0.003	-0.078**	-0.015
	(0.012)	(0.005)	(0.037)	(0.016)
Age	-0.040	-0.011**	-0.071***	-0.006
	(0.025)	(0.005)	(0.028)	(0.007)
Ethnic minority	-0.150*	0.013	-0.109	0.026
	(0.089)	(0.018)	(0.099)	(0.023)
Holds agricultural hukou	-0.057*	-0.010	-0.112*	-0.042
	(0.032)	(0.013)	(0.057)	(0.032)
Mother's education level				
Middle school	0.125***	0.009	0.140***	-0.008
	(0.031)	(0.013)	(0.046)	(0.022)
High/technical school	0.112***	0.003	0.115	0.043
	(0.035)	(0.013)	(0.074)	(0.038)
College or above	0.139***	0.005	0.066	-0.069
	(0.041)	(0.015)	(0.102)	(0.065)
Father's education level				
Middle school	0.038*	-0.010	0.065	-0.012
	(0.022)	(0.009)	(0.040)	(0.027)
High/technical school	0.022	0.000	0.018	-0.041
	(0.030)	(0.014)	(0.063)	(0.045)
College or above	0.051	0.010	0.267***	0.149***
	(0.036)	(0.017)	(0.075)	(0.061)
Low perceived ability in math	-0.058*	-0.015		
	(0.033)	(0.018)		
Grade-by-school fixed effects		X		X
Number of observations	8,294	8,294	850	850
R-squared	0.08	0.66	0.18	0.85
Joint test F-statistic	3.21	0.97	14.27	1.55
[p-value]	[0.00]	[0.48]	[0.00]	[0.12]

Notes: This table shows results from four separate regressions of math teacher gender (=1 if female) on the set of independent variables listed in the first column, following the test in Antecol et al. (2015). Columns 1 and 2 show estimates generated using the entire sample and columns 3 and 4 show estimates generated using the low perceived ability group only. * p<0.1, ** p<0.05, *** p<0.01.

grade of primary school, and we use this question to proxy for students' perception of their ability[14]. Specifically, we classify those students who found learning math in the sixth grade to be "very difficult" as low perceived ability. We classify those who report sixth grade math to be "somewhat difficult," "not so difficult," or "easy" not to be of low perceived ability. In Table A.3 we show characteristics of students, by gender, for both of the perceived ability groups. Gaps between boys and girls described earlier persist across groups though, consistent with stereotypes, a higher proportion of girls perceive themselves to be low perceived ability than do boys (11.7% vs. 8.9%).

It is important to note that this data is observed at the same time as all of the other data, specifically, after teacher assignment. It is possible, therefore, that teacher assignment may affect recall of perceived ability. To test for this possibility, we run the same regressions of teacher gender on predetermined characteristics, only restricting our analysis to low perceived ability students. We show our results in columns 3 and 4 of Table 1. The general pattern is the same as that for the entire sample - once we control for grade-by-school fixed effects, we fail to reject the null that these characteristics are jointly insignificant.

It is also possible that the determinants of perceived ability differ between boys and girls in a way that may predict their test scores. If this were true, it would influence our interpretation of γ_3. To examine this possibility, we regress test scores on the vector of student-level predetermined characteristics and, using these coefficients, generate a predicted test score for each student. 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 test scores between genders in either group.

As a robustness check and further safeguard against possible omitted variables bias in the perceived ability measure, we also generate estimates from an alternative specification. In this specification, we estimate the effect of teacher-student gender match on our main outcome variables (perceived difficulty of current math class, aspirations, stereo-

[14]This is not intended to proxy for a student's actual ability, but rather, to (noisily) measure how able she thinks herself to be.

typical beliefs, and math test scores) replacing the interaction terms for low perceived ability with an interaction for those whose math test score is below the median value of their teacher-gender pairing group (e.g., those boys paired with a male teacher whose test score is below the median score for that group). The results, presented in a table and figure in the appendix, remain largely similar in magnitude and significance.

5 The effects of being assigned a female math teacher

In this section, we present results from applying our empirical strategy to the CEPS data. First, we estimate the impact of teacher-student gender match in mathematics on student beliefs and aspirations. We then look at how this match affects investment in human capital, followed by analysis of its impact on performance on math tests.

5.1 Beliefs and aspirations

In this subsection, we conduct a test of the model's prediction that being assigned a female math teacher should positively affect beliefs and aspirations for low perceived ability girls. We investigate the impact of teacher gender on three belief variables: perceived difficulty of current math class, the careers to which students aspire, and anti-girl stereotypes. Our specification follows Equation 3, using grade-by-school fixed effects and the full battery of controls for students and teachers. We also control for students' math test scores, allowing us to compare changes in beliefs while controlling for performance.

For the analysis of perceived difficulty, we use the present-time analogue to the baseline perceived ability question[15], students' response to the prompt "how difficult do you find your current math class to be?" Potential responses were as before ("very difficult," "somewhat difficult," "not so difficult," and "not difficult at all") and we code the variable, as we do with perceived ability, with an indicator equal to one if the response is "very difficult."

[15]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 study the impact of teacher-student gender match on aspirations, we use children's response to the prompt "what job would you most like to do when you grow up?" There are several possible responses to the question[16]; our theory predicts the clearest break between jobs in the language and visual arts (designer; artist/actor), which are traditionally more common jobs for women in China, and everything else. For stereotypes, we estimate the effect of teacher-student gender match on whether the student agrees with a statement that boys are better than girls at learning math.

We present our results in Table 2. In column 1, our results suggest that being taught by a female math teacher reduces low perceived ability girls' probability of perceiving math as "very difficult" by 20 percentage points (γ_3). While the estimated effect for not-low perceived ability girls assigned to a female teacher is the same sign as for the low perceived ability girls, it is an order of magnitude smaller and not statistically significant (β_3). Being assigned a female math teacher is also associated with an increase in low perceived ability boys' perceived difficulty of math (γ_2). These results accord with predictions 1-3 from the model.

We also present results for low perceived ability students visually in Figure 3. In this figure, we plot the distribution of perceived difficulty of the current math class for each possible teacher-student gender pairing, restricting the sample to low perceived ability children. This shows the same pattern as the coefficients - girls assigned to a female teacher are at least 20 percentage points less likely to perceive math to be very difficult than any other group, and boys assigned to female math teachers are at least 10 percentage points more likely to find math very difficult than any other group. In column 1 of Table A.5, we present this estimate using students below the within-group median test score instead of the low perceived ability group. We observe below-median girls assigned to a female teacher are 7.8 percentage points less likely to find math very difficult. While smaller than our coefficient for the low perceived ability group, this estimate retains both the predicted sign and statistical significance. Figure A.3 gives the below-median analogue to Figure 3.

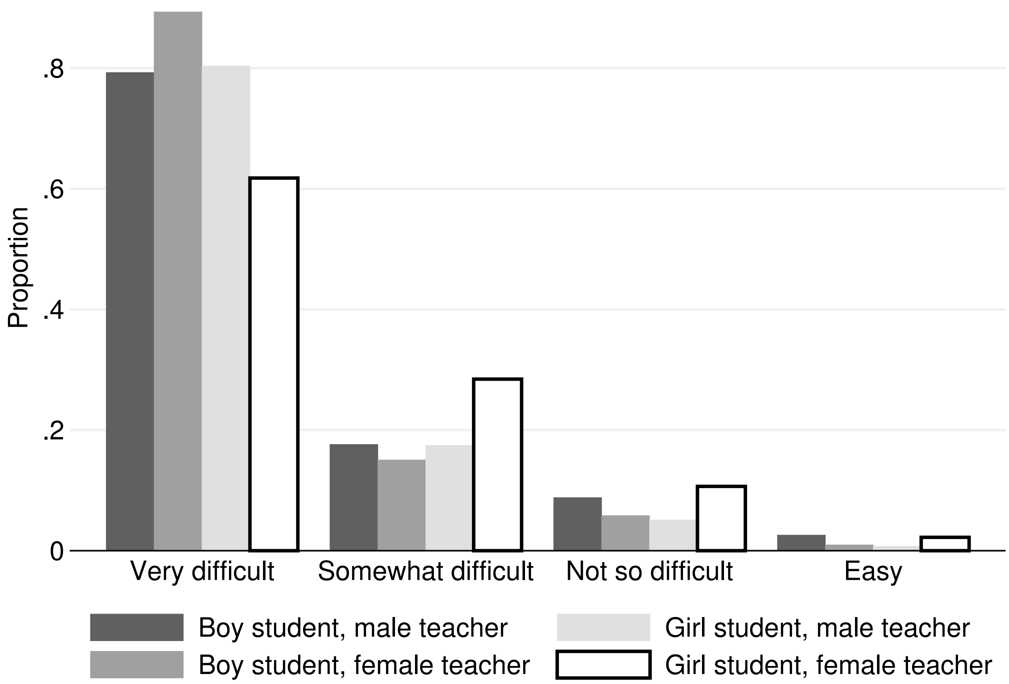
[16] The options are 1. Government Official, 2. Business manager, 3. Scientist/engineer, 4. Teacher/doctor/lawyer, 5. Designer, 6. Artist/actor, 7. Athlete, 8. Skilled worker, 9. Other, 10. Don't care, 11. Don't know.

Table 2: Effects on beliefs and aspirations

	(1) Current math class perceived as very difficult	(2) Aspires to jobs in art and design	(3) Holds anti-girl stereotypes
Girl x female teacher x low perceived ability	-0.205***	-0.110**	-0.038
	(0.057)	(0.056)	(0.070)
Female teacher x low perceived ability	0.100**	-0.031	0.079
	(0.046)	(0.034)	(0.058)
Girl x female teacher	-0.037	0.008	-0.047
	(0.017)	(0.019)	(0.035)
Girl x low perceived ability	0.046	0.105***	0.351***
	(0.042)	(0.034)	(0.053)
Girl	0.051***	0.184***	-0.130***
	(0.013)	(0.017)	(0.030)
Female teacher	0.010	0.004	0.045
	(0.017)	(0.018)	(0.032)
Low perceived ability	0.451***	-0.003	-0.196***
	(0.037)	(0.026)	(0.044)
Mean for non-LPA boys	0.122	0.104	0.599
Number of observations	8,276	8,213	8,117

Notes: The regression specification used is given in Equation 3, adding a control for the student's math test scores. Point estimates and their precision are largely unchanged by removing this final control. Robust standard errors clustered at the school level are shown in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Figure 3: Low perceived ability students' current perception of the difficulty of math, by gender of student and math teacher



Notes: This figure plots the response of low perceived ability students to the prompt: "how difficult do you find your current mathematics course to be?" This shows a clear rightward shift (towards lower perceived levels of difficulty in mathematics) for low perceived ability girls assigned to a female teacher, relative to all other teacher-student gender pairings.

In column 2 of Table 2, we present estimates of the effect of being assigned a female math teacher on students' career aspirations. We estimate that for low perceived ability girls, being assigned a female math teacher is associated with an 11 percentage point decrease in aspiring to traditionally female jobs. The positive and significant coefficients on the "girl" and "girl x low perceived ability" variables corroborate our choice of variable coding - girls, and particularly low perceived ability girls, are more likely to aspire to these jobs, independent of the gender of their teacher. The effects of being assigned a female teacher on all other groups (low perceived ability boys, all other boys and girls) are at least an order of magnitude smaller and insignificant. In Table A.5 we see no effect on aspirations for the below-median girls assigned to female teachers.

In column 3, our estimate shows a small, insignificant effect on girls' stereotypical beliefs. The total effect on the gender gap, $\beta_3 + \gamma_3$, however, is 8.5 percentage points, or a 13.4% decrease in the proportion of low perceived ability girls in our sample who hold these beliefs. In column 3 of Table A.5, we see a coefficient on girl x female teacher x below median of 8.0 percentage points, significant at the 10% level, and a total effect of 11.7 percentage points, or 25% of the baseline proportion. These results lead to two conjectures: one, that it may be harder to change global beliefs (stereotypes) than local beliefs (perceptions of own ability, as proxied by perceived difficulty); and two, that those with somewhat higher perceived ability may be more prone to updating their global beliefs when presented with a positive role model. In the context of our model, the mapping from a child's prior to the size of her update, as in Figure A.1, may be less right skewed for stereotypes than for perceived ability.

5.2 Investment in human capital

We next conduct a series of tests of the model's prediction that teacher-student gender match should positively change investment behavior for low perceived ability girls matched with female math teachers. We test this using four different dependent variables: students' reported enrollment in math tutoring, their total hours in tutoring (including, but not only, math tutoring), their hours spent on homework, and their enrollment in math olympiad tutoring. We give these results in Table 3.

Estimates presented in column 1 suggest that for low perceived ability girls, teacher-student gender match is associated with a 9.1 percentage point increase in enrollment in math tutoring. Low perceived ability boys assigned to female teachers, on the other hand, spend substantially less time in tutoring than those assigned to male teachers. These estimates of γ_2 and γ_3 also agree with our model's predictions, though it is worth noting that the total effect of being assigned a female teacher for low perceived ability girls, $\gamma_2 + \gamma_3$, is 0.9, meaning that they are on par with low perceived ability students assigned to male teachers. In Section 6, we explore the possible sources (i.e., parents, teachers, or children) of this change in investment. Also, and again as predicted, we see a much smaller and statistically insignificant estimate of β_3, the girl x female teacher coefficient.

In columns 2 and 3, we present estimates of the effect of teacher-student gender match on time-use data, first for weekly hours spent in tutoring, then for hours per week spent on homework. The pattern of estimates in column 2 is similar to that in column 1 - when assigned to a female math teacher, low perceived ability girls spend more time in tutoring (three hours per week), and low perceived ability boys spend slightly less (1.5 hours). These results are only suggestive, however, as the time use data is not specifically about math tutoring, but rather time spent in tutoring overall. In column 3, we see no significant effect on hours spent on homework for either group.

In column 4, we estimate the effect of being assigned a female math teacher on enrollment in math olympiad tutoring. This tutoring is designed for students who aim to develop advanced math skills. Since the low perceived ability girls also have lower math test scores than their peers, it is unlikely that the differences in beliefs apparently induced by a female math teacher would lead to substantial gains in olympiad tutoring, which is targeted at students of relatively higher ability. On the other hand, if role models also affect beliefs and behavior at the higher end of the perceived ability spectrum, we may find an impact on olympiad tutoring for higher perceived ability girls. In line with what our model predicts, we see no significant effect of being assigned a female math teacher on enrollment in math olympiad tutoring among girls at any perceived ability level. Finally, consistent with the posited negative influence of anti-girl stereotypes in math, girls in our sample are 30% (1.8 percentage points) less likely to enroll in math olympiad tutoring than boys despite girls' superior performance on mathematics examinations. We explore this gap in test scores further in the next section.

Table 3: Effects on investment in human capital

	(1) Enrolled in math tutoring	(2) Hours in tutoring	(3) Hours spent on homework	(4) Math olympiad tutoring
Girl x female teacher x low perceived ability	0.091*	3.057***	0.392	0.000
	(0.052)	(1.253)	(1.595)	(0.024)
Female teacher x low perceived ability	-0.082**	-1.548	0.687	-0.014
	(0.036)	(0.996)	(1.353)	(0.022)
Girl x female teacher	0.027	0.262	0.516	-0.006
	(0.019)	(0.363)	(0.519)	(0.011)
Girl x low perceived ability	-0.054	-2.203***	-1.044	-0.001
	(0.035)	(0.933)	(1.252)	(0.018)
Girl	0.022	0.080	0.716*	-0.018**
	(0.016)	(0.295)	(0.403)	(0.008)
Female teacher	-0.012	-0.262	0.095	0.018
	(0.023)	(0.403)	(0.467)	(0.016)
Low perceived ability	0.041	1.231	0.081	0.008
	(0.026)	(0.832)	(1.052)	(0.016)
Mean for non-LPA boys	0.210	4.046	5.545	0.063
Number of observations	8,257	8,019	7,995	8,257

Notes: The regression specification used here is given in Equation 3. The dependent variable is given in the column headings. Robust standard errors clustered at the school level are shown in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

5.3 Academic performance

In this subsection we examine the effect of teacher-student gender match on students' performance in mathematics. Here we focus on children's scores on midterm math examinations to quantify the differences apparent in Figure 2. We present these results in Table 4. The first column shows the estimates with no fixed effects and the second column shows estimates generated with grade-by-school fixed effects (the specification used Tables 2 and 3).

We find that having a female math teacher increases the math test scores of low perceived ability girls by 4.46 points, or 0.446 sample standard deviations (SDs), controlling for other characteristics as in Equation 3. In line with the predictions of the model, girls who do not perceive themselves of low ability appear to gain no gender-specific benefit from being assigned a female teacher ($\beta_3 = 0.068$, $\sigma = 0.541$). Consistent with the patterns shown in the previous subsections, we also see some evidence that low perceived ability boys' test scores decline, though the effect is not significant and is much smaller than the effect for low perceived ability girls.

While our estimate of 0.446 SD is quite large, it is estimated for a subgroup that our model predicts is particularly likely to benefit from teacher-student gender match. If we remove the LPA term and its interactions (i.e., all of the variables with γ coefficients), our estimate of β_3, the coefficient on teacher-student gender match, is 0.093 SD ($\sigma = 0.063$). This is well within the range of estimates generated in previous work (e.g., Dee, 2007; Muralidharan and Sundararaman, 2011; Lim and Meer, Forthcoming). Other work evaluating interventions in developing countries that targeted either low performers or those in particularly needy regions finds similarly large effects (Banerjee and Duflo, 2007; Burde and Linden, 2013).

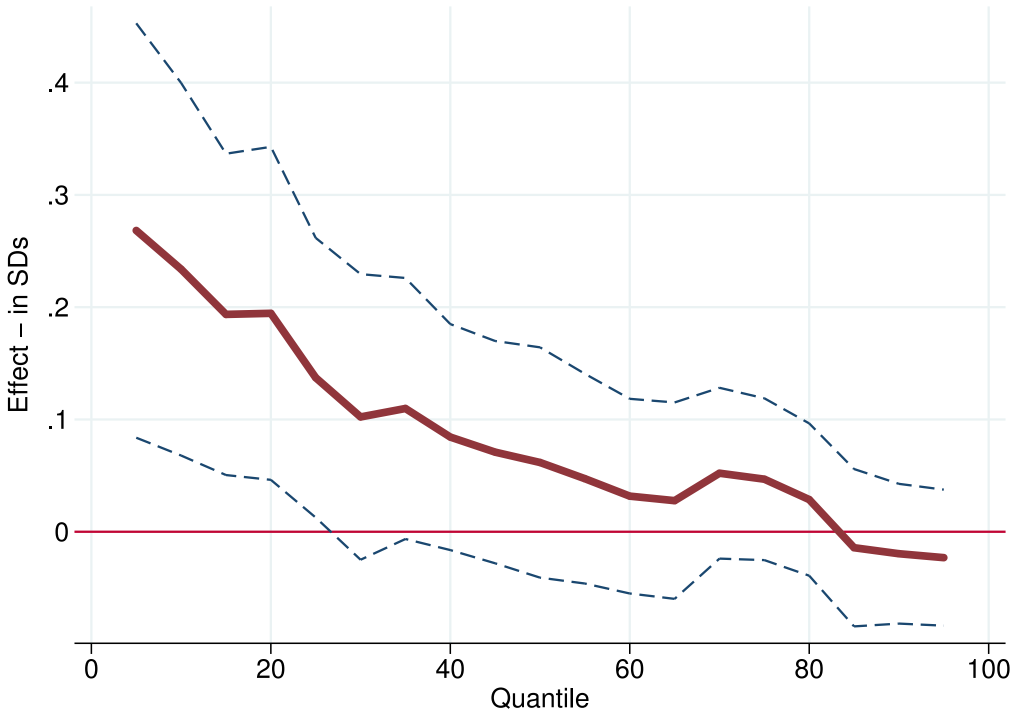
We present quantile regression results in Figure 4. To generate this figure, we estimate equation 3 without any of the low perceived ability independent variables and recover coefficient estimates of β_3 and the corresponding confidence interval at every fifth centile between the fifth and 95th. The pattern that we see in Table 4 and in Figure 2 also appears

Table 4: Effects on math test score

	(1)	(2)
Girl x female teacher	4.783***	4.459***
x low perceived ability	(1.621)	(1.657)
Female teacher	-2.294*	-1.472
x low perceived ability	(1.240)	(1.288)
Girl x female teacher	0.187	0.068
	(0.535)	(0.541)
Girl x low perceived ability	-0.414	-0.193
	(1.267)	(1.248)
Girl	1.191***	1.250***
	(0.482)	(0.492)
Female teacher	0.406	1.849***
	(0.423)	(0.684)
Low perceived ability	-7.252***	-8.064***
	(0.817)	(0.844)
Mean for non-LPA boys	70.242	70.242
Number of observations	8,294	8,294
Grade-by-school fixed effects		X

Notes: The dependent variable is the student's math test score. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in Equation 3. * p<0.1, ** p<0.05, *** p<0.01.

Figure 4: Quantile regression results for math test scores



Note: This figure presents coefficient estimates and standard errors of β_3, estimated at every fifth quantile from the fifth to 95th, using Equation 3 but removing the low perceived ability controls and their interactions (i.e., all of the terms with γ coefficients). The dependent variable is midterm math test score.

in Figure 4 - the gains from teacher-student gender match accrue to those girls in the left half of the distribution, and the largest gains accrue to those in the first quartile. In column 4 of Table A.5, we estimate a positive effect of teacher-student gender match on math test scores for the below-median group of girls assigned to female math teachers. These results are substantially smaller than our estimates generated using perceived ability (around 0.11 SDs for the below-median group, as opposed to the 0.45 SD gain we measure for the low perceived ability group) and for the left quartile of the distribution in the quantile regressions (0.2-0.3 SD). This discrepancy is also in line with our model's predictions. While the below-median and quartile results are both sizable, the framework in Section 2 predicts that it is specifically among the low perceived ability girls, not just the low performers, that we should see the largest difference.

6 Mechanisms and discussion

In this section, we first conduct a series of analyses to test for evidence of two potential drivers of the patterns we observe in the previous section - the first, described in our conceptual framework, is that same-gendered math teachers serve as role models who counter the negative effects of stereotypes on student beliefs; the second is that there is some other characteristic of female teachers or their conduct which drives these results. We then provide a discussion of the main limitations of our analysis.

6.1 Mechanisms

In this subsection we first show that additional exposure to female role models in mathematics has additional positive effects on the outcomes we study. We then perform a series of analyses looking for evidence that the effects we observe in the previous section are driven by either teacher-specific characteristics or teacher conduct instead of by the effect of the teacher serving as a role model. We test for the following possibilities: one, that female math teachers lavish more attention on low perceived ability girls than do male teachers; two, that female math teachers are merely better teachers and it is these skill differentials which drive the observed effects; three, that female teachers exerting more effort than male teachers drives the effects; and four, that our findings are driven by differences in teaching methods between female and male teachers.

We first show that additional exposure to the role model increases the impact of teacher-student gender match on the outcomes we measure. We exploit the fact that some students' math teacher is also their homeroom teacher, and these students spend additional time each day in the presence of that teacher. This provides variation in the amount of students' exposure to the role model. Our framework predicts that this additional exposure should generate additional positive effects, and we test this by estimating the effect of a student being assigned a female math teacher who is also the student's homeroom teacher on beliefs, stereotypes, enrollment in math tutoring, and performance on the midterm math exam. We present these results in Table A.6. While the results are imprecise, the coefficients are large and, for all but the tutoring variable, in the predicted direction.

Next, we present results from a series of tests that consider alternative explanations for the patterns we observe in Section 5. We test four possibilities: one, that female math teachers may choose to engage more with low perceived ability girls than do male math teachers; two, that female math teachers may simply be better at teaching those of low perceived ability than male teachers; three, that it is differential teacher effort instead which drives the results we observe; and four, that female teachers teach differently than do male teachers, and this difference is responsible for the salutary effects of teacher-student gender match that we measure.

First we investigate whether female teachers favor girls with more praise and attention. (Beaman et al., 2009; Hoffmann and Oreopoulos, 2009; Jones and Wheatley, 1990). The CEPS collects students' recall of how frequently their current math teacher asks them questions and their recall of how frequently the teacher praises them in the classroom. In Table 5 we present results from estimating Equation 3 using these two measures as outcome variables[17]. Our results show that while female teachers are slightly more likely to ask students questions than male teachers, there is no evidence that female teachers favor low perceived ability girls either with more opportunities to respond to questions or more praise.

The second possibility is that female 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

[17] 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 5: 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	0.035	-0.083
	(0.080)	(0.069)
Female teacher x low perceived ability	-0.084	-0.022
	(0.055)	(0.043)
Girl x female teacher	0.008	0.030
	(0.025)	(0.024)
Girl x low perceived ability	0.012	0.022
	(0.053)	(0.048)
Girl	-0.030	-0.055
	(0.022)	(0.020)
Female teacher	0.057	0.024
	(0.035)	(0.036)
Low perceived ability	-0.077*	-0.097
	(0.043)	(0.030)
Mean for non-LPA boys	0.635	0.513
Number of observations	8,237	8,246

Notes: The regression specification used here is given in Equation 3, again with the addition of the midterm math test score. 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. All regressions control for the student's math test scores, but the point estimates and their precision are largely unchanged by removing this control. * p<0.1, ** p<0.05, *** p<0.01.

term for girl x math teacher who won an award. We show these results in Table A.7. These results do not show no evidence of “better” teachers having a positive effect on perceived difficulty, aspirations, or performance of low perceived ability girls, though they appear to affect stereotypical beliefs. To probe this further, we also conduct a horse race, reverting to the original specification in Equation 3 and adding the interaction variable for award-winning teacher with the teacher-student gender match dummy. 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[18]. We generate (but do not show) similar results for two separate specifications swapping the receipt of a teaching award with either one, years of experience, or two, holding a degree from a teacher training (normal) university.

Next, we investigate the possibility that teacher effort drives these effects. 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 the effort variable. 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 different groups of students. We estimate the effects of effort 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. In Tables A.8, A.9, and A.10, we show results for the same dependent variables: perceived difficulty of math, whether the child holds anti-girl stereotypes, enrollment in math tutoring, and the math midterm exam score. For none of these analyses do we observe a significant effect of teacher effort on outcomes for low perceived

[18] Results in tabular form are available from the authors but not included in this manuscript.

ability girls.

Finally, we look at the impact of teachers' use of different methods of teaching on low perceived ability girls. 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) rather than the role model effect. The CEPS records teachers' response to the following question - "how often do you use [teaching method]: never, sometimes, often, or always?" - 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 and so we expect a priori for them to have a larger effect on the 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 Limitations

In this subsection, we outline a few limitations of our analysis, focusing on areas of departure from the ideal scenario in which to test our main hypotheses. Our first and main weakness is the lack of multiple time periods. Ideally we would observe children's performance in primary school, then observe their beliefs when they are paired with their teacher and estimate those effects, then observe investment decisions and, finally, observe performance. Instead, we observe a series of demographic characteristics and data collected in one period, usually in the student's first year of middle school. It is possible, therefore, that the arrow of causality goes the other way, from some unobserved factor that boosts performance for these girls, which in turn changes perceived difficulty, investment, and aspirations. While possible, it would have to be the case that this unobserved factor accrues only to low perceived ability girls in math classes. In the previous subsection we presented a series of results suggesting that neither teacher aptitude nor any of a battery of teacher behaviors are likely to be the source of such reverse causality.

The second weakness is the use of the self-reported perceived ability measure. In an ideal setting we would observe children's actual performance in the sixth grade, and use this as a stratifying variable. While we have shown no evidence of difference in observable characteristics associated with this variable that would suggest omitted variables bias, we cannot definitively exclude the possibility of some unobserved factor which affects both propensity to report low perceived ability and our main outcome variables. Nonetheless, the fact that we see similar patterns in perceived difficulty of math and math performance for both the below-median group and the low perceived ability group suggests that our conceptual framework's main predictions for the signs of our estimates appear to be borne out.

This study looks at the effects of teacher-student gender match in mathematics, a subject where girls face longstanding stereotypes against their ability. A good ancillary test of our theory would be to test for effects of teacher-student gender match on beliefs and test performance in subjects without stereotypes, namely, in English and Chinese. While we have information on teachers in these subjects, the very small number of male English or Chinese teachers in our data prevents us from using our identification strategy, which clusters at the grade-by-school level, to test for such effects. We present the distribution of Chinese and English test scores by teacher-student gender pairing in Figure A.4, and note that the large difference seen in the math scores of low perceived ability girls assigned to female teachers is not apparent; overall, we see some evidence of (substantially smaller) effects of teacher-student gender match on beliefs and test scores. Because of the large standard errors and the reduced sample size - we have to exclude grades in schools without at least one male Chinese or English teacher - we are unable to reject a zero effect.

Finally, we observe a change in children's enrollment in tutoring which admits several possible explanations. One possible explanation at odds with our interpretation of the results is that parents' and/or teachers' compensatory actions, including but not limited to enrolling low perceived ability children in more tutoring, causes the change in beliefs, aspirations, and performance we observe. While we do not deny the possibility that this may play some role for some students, the patterns in our empirical results and a few facts about the Chinese context suggest this is similarly unlikely to be the most important driver of our empirical results. First, our analyses of teacher effort and interaction with students by gender show no evidence of differential teacher attention or effort driving the differences we observe. Second, were compensatory behavior by parents to drive this pattern, it would have to be the case that parents of boys assigned to female teachers respond by withdrawing their children from tutoring while the parents of girls assigned to female teachers respond by increasing enrollment in tutoring. Our explanation - these results come from a difference in enthusiasm, effort, and belief in oneself generated by the role model effect of being assigned a same-gendered teacher - is rooted in existing empirical and theoretical evidence from both economics and psychology (e.g., Bettinger and Long, 2005; Nixon and Robinson, 1999; Paredes, 2014; Lybbert and Wydick, 2016b; Bian et al., 2017). We argue that this explanation is more plausible and less of a "just-so" story than subgroup-specific compensatory action by the parents of low perceived perceived ability girls and boys assigned to female teachers. Lastly, we study a context where children are often actively involved in their education, particularly in the age range we study. Children in the CEPS are in middle school when our data is collected. Loyalka et al. (2013) find that an information intervention providing students in a different set of Chinese middle schools with estimated labour market returns to different levels of education affected these students' propensity to drop out of middle school. This evidence is consistent with the notion that children in Chinese middle schools make at least some of their own educational decisions.

7 Conclusion

In this paper, we study how stereotypes and role models affect the formation of children's beliefs about their ability. We model this process and take our model's predictions to an empirical context where strong negative stereotypes regarding girls' math ability vis-a-vis boys' prevail despite girls' superior performance in math. We find that low perceived ability students benefit from being assigned a same-gendered teacher, and these benefits are largest for students who face stereotypes (girls) and are on the margin of giving up on themselves, and has little effect for students, regardless of gender, who are well above this margin.

Our paper generates two broad messages. First, it shows that even in the increasingly common case of a reverse gender gap, i.e., where girls outperform boys, anti-girl stereotypes can persist and role models may ameliorate their negative effects on the most vulnerable. Second, it adds to a growing body of research (e.g., Bernard et al., 2014; Lusher et al., 2015; Lybbert and Wydick, 2016b; Genicot and Ray, 2017) in economics studying how information affects aspirations, educational decisions and outcomes. Together, this work shows that the informational environment a child faces and, specifically, the presence of a plausible example of success, may be a key lever for changing beliefs, increasing effort, and improving performance in school. More broadly, our work suggests that role models and the information they provide are likely an important input into the production of human capital, particularly among girls and other groups who for various historical or socioeconomic reasons may lack for a credible example of successful investment in certain types of human capital.

References

Akerlof, G.A. and R.E. Kranton, "Identity and schooling: Some lessons for the economics of education," *Journal of Economic Literature*, 2002, *40* (4), 1167–1201.

Akerlof, George A. and Rachel E. Kranton, "Economics and Identity," *Quarterly Journal of Economics*, 2000, *115* (3), 715–753.

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.

_ **, Shawn Cole, Esther Duflo, and Leigh Linden**, "Remedying education: Evidence from two randomized experiments in India," *Quarterly Journal of Economics*, 2007, *122* (3), 1235–1264.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, Petia Topalova et al., "Powerful Women: Does Exposure Reduce Bias?," *Quarterly Journal of Economics*, 2009, *124* (4), 1497–1540.

Becker, Gary S, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, National Bureau of Economic Research, Inc, 1975.

Beilock, Sian L, Elizabeth A Gunderson, Gerardo Ramirez, and Susan C Levine, "Female teachers' math anxiety affects girls' math achievement," *Proceedings of the National Academy of Sciences*, 2010, *107* (5), 1860–1863.

Bénabou, Roland and Jean Tirole, "Identity, morals, and taboos: Beliefs as assets," *Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

Bernard, Tanguy, Stefan Dercon, Kate Orkin, and Alemayehu Taffesse, "The future in mind: aspirations and forward-looking behaviour in rural Ethiopia," *BREAD Working Paper*, 2014, (429).

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.

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.

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.

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

Fairlie, Robert W, Florian Hoffmann, and Philip Oreopoulos, "A community college instructor like me: Race and ethnicity interactions in the classroom," *American Economic Review*, 2014, *104* (8), 2567–2591.

Genicot, Garance and Debraj Ray, "Aspirations and Inequality," *Econometrica*, 2017, *85* (2), 489–519.

Gershenson, Seth, 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*, Forthcoming.

Gunderson, Elizabeth A, Gerardo Ramirez, Susan C Levine, and Sian L Beilock, "The role of parents and teachers in the development of gender-related math attitudes," *Sex Roles*, 2012, *66* (3-4), 153–166.

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.

Jensen, Robert, “The (perceived) returns to education and the demand for schooling,” *Quarterly Journal of Economics*, 2010, *125* (2), 515–548.

Jones, M Gail and Jack Wheatley, “Gender differences in teacher-student interactions in science classrooms,” *Journal of Research in Science Teaching*, 1990, *27* (9), 861–874.

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.

Lai, Fang, Elisabeth Sadoulet, and Alain De Janvry, “The contributions of school quality and teacher qualifications to student performance evidence from a natural experiment in Beijing middle schools,” *Journal of Human Resources*, 2011, *46* (1), 123–153.

Latane, Bibb, “The psychology of social impact.,” *American Psychologist*, 1981, *36* (4), 343.

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland, “Expectations of brilliance underlie gender distributions across academic disciplines,” *Science*, 2015, *347* (6219), 262–265.

Lim, Jaegeum and Jonathan Meer, “The impact of teacher-student gender matches: random assignment evidence from South Korea,” *Journal of Human Resources*, Forthcoming.

Loyalka, Prashant, Chengfang Liu, Yingquan Song, Hongmei Yi, Xiaoting Huang, Jianguo Wei, Linxiu Zhang, Yaojiang Shi, James Chu, and Scott Rozelle, “Can information and counseling help students from poor rural areas go to high school? Evidence from China," *Journal of Comparative Economics*, 2013, *41* (4), 1012–1025.

Lusher, Lester, Doug Campbell, and Scott Carrell, "TAs like me: racial interactions between graduate teaching assistants and undergraduates," *National Bureau of Economic Research*, 2015, (w21568).

Lybbert, Travis J and Bruce Wydick, "Hope as aspirations, agency, and pathways: poverty dynamics and microfinance in Oaxaca, Mexico," *National Bureau of Economic Research Working Paper*, 2016, (w22661).

_ **and** _ , "Poverty, aspirations, and the economics of hope," *University of California, Davis Working Paper*, 2016.

Muralidharan, Karthik 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.

Nixon, Lucia A and Michael D Robinson, "The educational attainment of young women: Role model effects of female high school faculty," *Demography*, 1999, *36* (2), 185–194.

OECD, *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*, OECD Publishing, 2015.

Paredes, Valentina, "A teacher like me or a student like me? Role model versus teacher bias effect," *Economics of Education Review*, 2014, *39*, 38–49.

Ross, Phillip H, "Aspirations and human capital investment: evidence from Indian adolescents," *Working Paper*, 2016.

Scheepers, Daan and Naomi Ellemers, "When the pressure is up: The assessment of social identity threat in low and high status groups," *Journal of Experimental Social Psychology*, 2005, *41* (2), 192–200.

Sherman, David K, Kimberly A Hartson, Kevin R Binning, Valerie Purdie-Vaughns, Julio Garcia, Suzanne Taborsky-Barba, Sarah Tomassetti, A David Nussbaum, and Geoffrey L Cohen, "Deflecting the trajectory and changing the narrative: how self-affirmation affects academic performance and motivation under identity threat.," *Journal of Personality and Social Psychology*, 2013, *104* (4), 591–618.

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.

Steele, Claude M, Steven J Spencer, and Joshua Aronson, "Contending with group image: The psychology of stereotype and social identity threat," *Advances in Experimental Social Psychology*, 2002, *34*, 379–440.

Steele, C.M. and J. Aronson, "Stereotype threat and the intellectual test performance of African Americans.," *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, 1995, *69* (5), 797.

Steele, Jennifer, "Children's gender stereotypes about math: The role of stereotype stratification," *Journal of Applied Social Psychology*, 2003, *33* (12), 2587–2606.

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

Zhang, Hongliang, “Identification of treatment effects under imperfect matching with an application to Chinese elite schools,” *Journal of Public Economics*, 2016, *142.*

Appendix

Appendix A: Appendix tables

Table A.1: Summary statistics for students

	(1) All	(2) Female	(3) Male	(4) Difference
Female (%)	48.71	-	-	-
Age	13.22	13.16	13.27	-0.10***
Minority (%)	11.31	11.78	10.86	0.92
Agricultural hukou (%)	48.44	47.55	49.28	-1.72
Father's years of education	10.69	10.75	10.62	0.13*
Father's highest level of schooling (%)				
Primary or below	13.77	13.46	14.07	-0.61
Middle school	41.14	40.93	41.33	-0.4
High school/technical school	25.43	25.04	25.79	-0.75
College or above	19.66	20.57	18.81	1.76**
Mother's years of education	9.97	10.08	9.87	0.20**
Mother's highest level of schooling (%)				
Primary or below	22.1	20.34	23.76	-3.42***
Middle school	38.11	39.7	36.59	3.12***
High school/technical school	22.92	22.58	23.25	-0.66
College or above	16.87	17.37	16.4	0.97
Number of siblings	0.69	0.75	0.64	0.11***
Household income "low"	18.11	16.97	19.18	-2.21***
Math test score	70.25	70.94	69.59	1.35***
Number of observations	8,345	4,065	4,280	-

Notes: Column 4 shows the gender differences in student characteristics with t-test results. * p<0.1, ** p<0.05, *** p<0.01.

Table A.2: Summary statistics for teachers

	(1) All	(2) Female	(3) Male	(4) Difference
Female (%)	61.35	-	-	-
Age	37.94	36.95	39.5	-2.55**
Education level (%)				
Associate college or below	12.56	7.87	20	-12.13**
Part-time four-year university	34.78	33.07	37.5	-4.43
Full-time four-year university	48.79	54.33	40	14.33**
Master's degree or higher	3.86	4.72	2.5	2.22
Attended a normal university (%)	94.2	92.13	97.5	-5.37
Years of teaching experience	16.8	15.72	18.53	-2.81**
Holds a senior professional rank (%)	23.67	24.41	22.5	1.91
Won teaching award (%)				
At the province or national level	14.01	14.96	12.5	2.46
At the city level	43.96	42.52	46.25	-3.73
Observations	207	127	80	-

Notes: This table compares observable teacher characteristics across teacher gender. Column 4 shows the gender differences in teacher characteristics with t-test results. * p<0.1, ** p<0.05, *** p<0.01.

Table A.3: Background characteristics, summarised by gender and perceived ability

	Perceived ability			
	Low		*Not low*	
	(1) Girls	(2) Boys	(3) Girls	(4) 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

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.

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

	(1)	(2)
Age	-0.010	-0.018
	0.009	0.030
Has B.A.	0.122	0.055
	0.078	0.249
Went to teachers' college	-0.242*	-0.222
	0.131	0.216
Years of experience	0.001	0.015
	0.008	0.027
Won award at province level	0.099	0.161
	0.115	0.387
Won award at city level	-0.027	-0.108
	0.073	0.255
Grade-by-school fixed effects		X
Number of observations	207	207
R-squared	0.06	0.70
Joint test F-statistic	2.31	0.25
[p-value]	[0.04]	[0.96]

Notes: This table shows coefficient and standard error estimates from regressing teacher gender on the predetermined teachers characteristics listed in the first column and conducting a Wald Test for their joint significance, similar to the results shown in Table 1 for student characteristics.

Table A.5: Replicating main results, 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) Believes boys are better than girls at learning math	(4) Midterm math test score
Girl x female teacher x below median	-0.078***	0.009	-0.080*	1.111*
	(0.027)	(0.034)	(0.046)	(0.569)
Female teacher x below median	0.007	-0.004	0.045	0.873***
	(0.024)	(0.019)	(0.034)	(0.334)
Girl x female teacher	-0.039**	-0.013	-0.028	0.116
	(0.017)	(0.022)	(0.041)	(0.289)
Girl x below median	0.068***	0.026	0.287***	0.558
	(0.022)	(0.026)	(0.034)	(0.462)
Girl	0.058***	0.188***	-0.224***	0.468**
	(0.014)	(0.019)	(0.034)	(0.236)
Female teacher	0.026	0.006	0.036	0.607**
	(0.020)	(0.018)	(0.039)	(0.303)
Low perceived ability	-0.031	-0.002	-0.112***	-16.531***
	(0.021)	(0.023)	(0.031)	(0.288)
Mean for non-LPA boys	0.069	0.085	0.638	77.571
Number of observations	8,300	8,251	8,151	8,345

Notes: The dependent variable in question is given in the column heading. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in Equation 3. * p<0.1, ** p<0.05, *** p<0.01.

Table A.6: Effect of having math teacher as homeroom teacher

	(1) Perceived difficulty of current math class	(2) Believes boys are better than girls at learning math	(3) Enrolled in after-school math tutoring	(4) Midterm math test score
HRMT x female teacher x girl x LPA	-0.113	-0.284**	-0.187	1.455
	(0.128)	(0.131)	(0.121)	(3.564)
HRMT x female x girl	0.027	0.053	-0.036	-0.166
	(0.034)	(0.078)	(0.047)	(1.117)
HRMT x LPA x girl	0.073	0.157	0.042	0.949
	(0.082)	(0.109)	(0.081)	(2.286)
Homeroom teacher = math teacher (HRMT) x LPA)	-0.033	-0.100	-0.068	-1.889
	(0.083)	(0.108)	(0.050)	(1.805)
HRMT x LPA	0.010	-0.003	0.169**	-1.863
	(0.102)	(0.132)	(0.078)	(2.582)
HRMT x female teacher	-0.084**	0.068	-0.048	3.468**
	(0.042)	(0.072)	(0.056)	(1.559)
HRMT x girl	-0.022	-0.087	0.014	-0.351
	(0.026)	(0.060)	(0.036)	(0.959)
Girl	0.057***	-0.100***	0.018	1.372***
	(0.017)	(0.034)	(0.019)	(0.553)
HRMT	0.008	0.022	0.041	-0.885
	(0.031)	(0.054)	(0.037)	(1.116)
Female math teacher	0.029	0.029	0.008	0.785
	(0.023)	(0.033)	(0.032)	(0.790)
Low perceived ability	0.464***	-0.161***	0.065*	-7.381***
	(0.038)	(0.055)	(0.034)	(0.896)
Mean for non-LPA boys	0.122	0.599	0.210	70.242
Number of observations	8,276	8,117	8,257	8,294

Notes: The dependent variable in question is given in the column heading. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in Equation 3. * p<0.1, ** p<0.05, *** p<0.01.

Table A.7: Teacher aptitude

	(1) Perceived difficulty of current math class	(2) Aspires to jobs in art and design	(3) Believes boys are better than girls at learning math	(4) Midterm math test score
Girl x award-winning teacher	-0.118	0.007	-0.194***	-1.174
x low perceived ability	(0.101)	(0.123)	(0.080)	(1.543)
Award-winning teacher	0.030	0.046	0.073	2.267*
x low perceived ability	(0.072)	(0.067)	(0.071)	(1.260)
Girl x award-winning teacher	-0.020	-0.039**	0.010	0.184
	(0.021)	(0.020)	(0.051)	(0.664)
Girl	-0.036	0.052*	0.359***	2.098**
x low perceived ability	(0.037)	(0.029)	(0.042)	(0.988)
Girl	0.031***	0.196***	-0.161***	1.275***
	(0.008)	(0.011)	(0.021)	(0.321)
Award-winning teacher	0.022	0.062**	-0.036	-0.301
	(0.025)	(0.029)	(0.041)	(0.952)
Low perceived ability	0.498***	-0.027	-0.163***	-9.143***
	(0.028)	(0.021)	(0.031)	(0.743)
Mean for non-LPA boys	0.122	0.104	0.599	70.242
Number of observations	8,276	8,213	8,117	8,294

Notes: The dependent variable in question is given in the column heading. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in Equation 3. * p<0.1, ** p<0.05, *** p<0.01.

Table A.8: Teacher effort 1

	(1) Perceived difficulty of current math class	(2) Believes boys are better than girls at learning math	(3) Enrolled in after-school math tutoring	(4) Midterm math test score
Hours prep: hours in class x low perceived ability x girl	0.037	-0.068*	0.021	-0.617
	(0.056)	(0.039)	(0.033)	(1.229)
Hours prep: hours in class x low perceived ability	-0.018	0.032	-0.006	1.516**
	(0.032)	(0.034)	(0.024)	(0.759)
Hours prep: hours in class x girl	-0.007	0.003	-0.011	0.511
	(0.007)	(0.015)	(0.012)	(0.323)
Girl x low perceived ability	-0.093	0.406***	-0.030	2.826
	(0.073)	(0.058)	(0.055)	(1.775)
Girl	0.036***	-0.161***	0.051***	0.748
	(0.011)	(0.026)	(0.018)	(0.485)
Hours prep: hours in class	0.019*	0.001	-0.033***	-0.972*
	(0.011)	(0.020)	(0.010)	(0.583)
Low perceived ability	0.520***	-0.192***	0.003	-10.552***
	(0.046)	(0.042)	(0.031)	(1.082)
Mean for non-LPA boys	0.122	0.599	0.210	70.242
Number of observations	8,212	8,055	8,193	8,230

Notes: The dependent variable in question is given in the column heading. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in Equation 3. * p<0.1, ** p<0.05, *** p<0.01.

Table A.9: Teacher effort 2

	(1) Perceived difficulty of current math class	(2) Believes boys are better than girls at learning math	(3) Enrolled in after-school math tutoring	(4) Midterm math test score
Hours grading: hours in class x low perceived ability x girl	0.013	-0.024	0.022	-0.990
	(0.042)	(0.041)	(0.050)	(1.299)
Hours grading: hours in class x low perceived ability	-0.053	0.015	0.000	1.488*
	(0.036)	(0.034)	(0.030)	(0.872)
Hours grading: hours in class x girl	-0.011	-0.011	0.006	0.512*
	(0.007)	(0.019)	(0.012)	(0.301)
Girl x low perceived ability	-0.071	0.357***	-0.031	3.192*
	(0.064)	(0.063)	(0.055)	(1.858)
Girl	0.041***	-0.145***	0.034**	0.751*
	(0.011)	(0.029)	(0.017)	(0.451)
Hours grading: hours in class	0.014	0.059***	-0.004	0.752
	(0.013)	(0.025)	(0.016)	(0.565)
Low perceived ability	0.559***	-0.171***	-0.004	-10.498***
	(0.051)	(0.045)	(0.034)	(1.222)
Mean for non-LPA boys	0.122	0.599	0.210	70.242
Number of observations	8,212	8,055	8,193	8,230

Notes: The dependent variable in question is given in the column heading. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in Equation 3. * p<0.1, ** p<0.05, *** p<0.01.

Table A.10: Teacher effort 3

	(1) Perceived difficulty of current math class	(2) Believes boys are better than girls at learning math	(3) Enrolled in after-school math tutoring	(4) Midterm math test score
Hours prep + grading: hours in class x low perceived ability x girl	0.016	-0.034	0.016	-0.535
	(0.032)	(0.024)	(0.022)	(0.760)
Hours prep + grading: hours in class x low perceived ability	-0.024	0.017	-0.002	1.053**
	(0.022)	(0.019)	(0.016)	(0.505)
Hours prep + grading: hours in class x girl	-0.006*	-0.003	-0.001	0.334*
	(0.004)	(0.009)	(0.007)	(0.178)
Girl x low perceived ability	-0.090	0.405***	-0.040	3.340
	(0.082)	(0.068)	(0.058)	(2.123)
Girl	0.042***	-0.152***	0.043**	0.573
	(0.012)	(0.029)	(0.019)	(0.502)
Hours prep + grading: hours in class	0.012*	0.017	-0.015*	-0.195
	(0.006)	(0.013)	(0.008)	(0.354)
Low perceived ability	0.554***	-0.194***	-0.000	-11.211***
	(0.058)	(0.046)	(0.036)	(1.343)
Mean for non-LPA boys	0.122	0.599	0.210	70.242
Number of observations	8,212	8,055	8,193	8,230

Notes: The dependent variable in question is given in the column heading. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in Equation 3. * p<0.1, ** p<0.05, *** p<0.01.

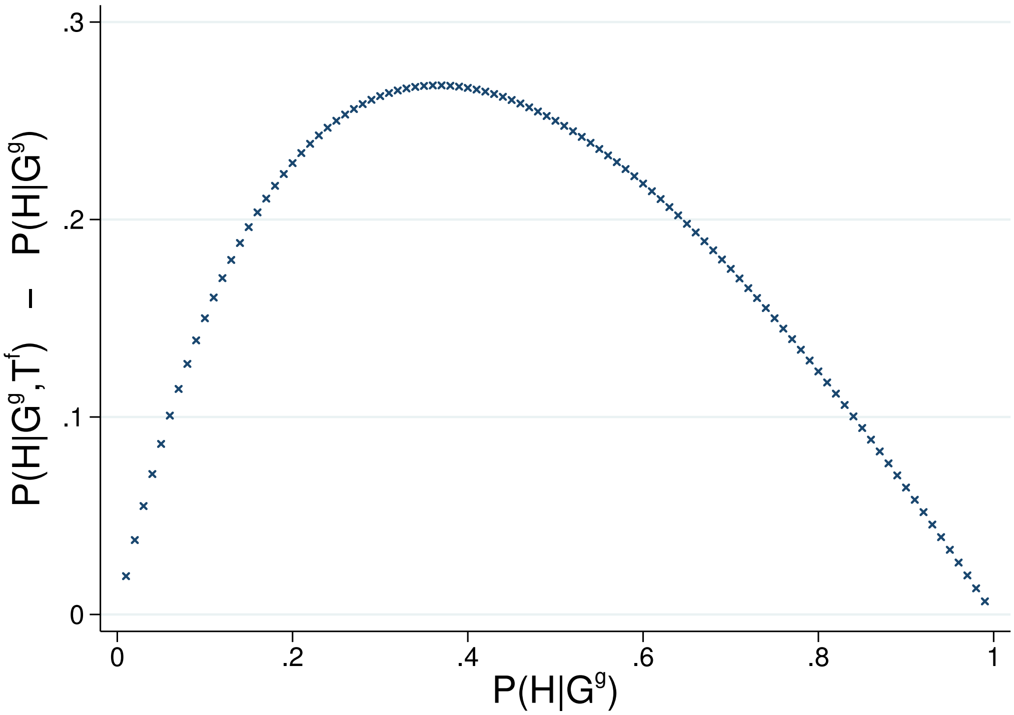
Table A.11: Teaching method

	Discuss in small groups		*Students and teacher "interactively" discuss*	
	(1) Perceived difficulty of current math class	(2) Midterm math test score	(3) Perceived difficulty of current math class	(4) Midterm math test score
Uses teaching method x girl x LPA	0.011	-1.035	0.048	-1.575
	(0.063)	(1.530)	(0.066)	(1.554)
Uses teaching method x LPA	0.028	0.011	-0.030	-1.149
	(0.048)	(1.232)	(0.050)	(1.313)
Uses teaching method x girl	-0.008	0.219	-0.033	-0.522
	(0.013)	(0.487)	(0.022)	(0.622)
Girl x LPA	-0.056	2.538**	-0.083*	3.105***
	(0.048)	(1.220)	(0.043)	(1.090)
Girl	0.031***	1.187***	0.053***	1.668***
	(0.011)	(0.330)	(0.020)	(0.536)
Uses teaching method	0.011	-0.456	0.007	-0.144
	(0.020)	(0.954)	(0.028)	(1.164)
Low perceived ability (LPA)	0.488***	-8.888***	0.520***	-8.007***
	(0.027)	(0.889)	(0.035)	(1.069)
Mean for non-LPA boys	0.122	70.242	0.122	70.242
Number of observations	8,257	8,275	8,251	8,268

Notes: The dependent variable in question is given in the column heading. Robust standard errors clustered at the school level are shown in parentheses, and the coefficients are estimated using the specification in Equation 3. * p<0.1, ** p<0.05, *** p<0.01.

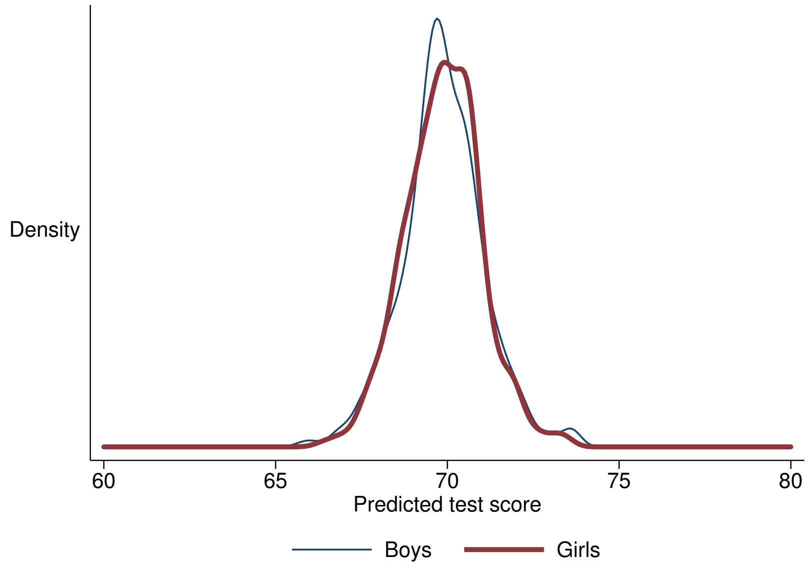
Appendix B: Appendix figures

Figure A.1: Mapping of prior to size of update

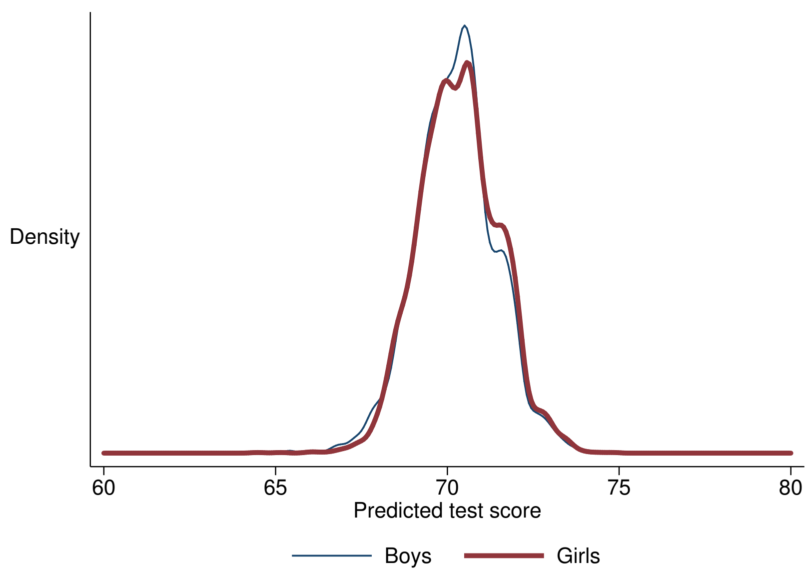


Notes: this figure shows the mapping from a girl's prior that she is of high ability, $P(H|G^g)$, to the update of that prior in response to encountering a female math teacher. The assumptions used to generate this figure are $P(G^g,T^f|H) = 0.6$ and $P(G^g,T^f|L) = 0.2$, but the right-skewness of the mapping generally holds under $P(G^g,T^f|H) > P(G^g,T^f|L)$.

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



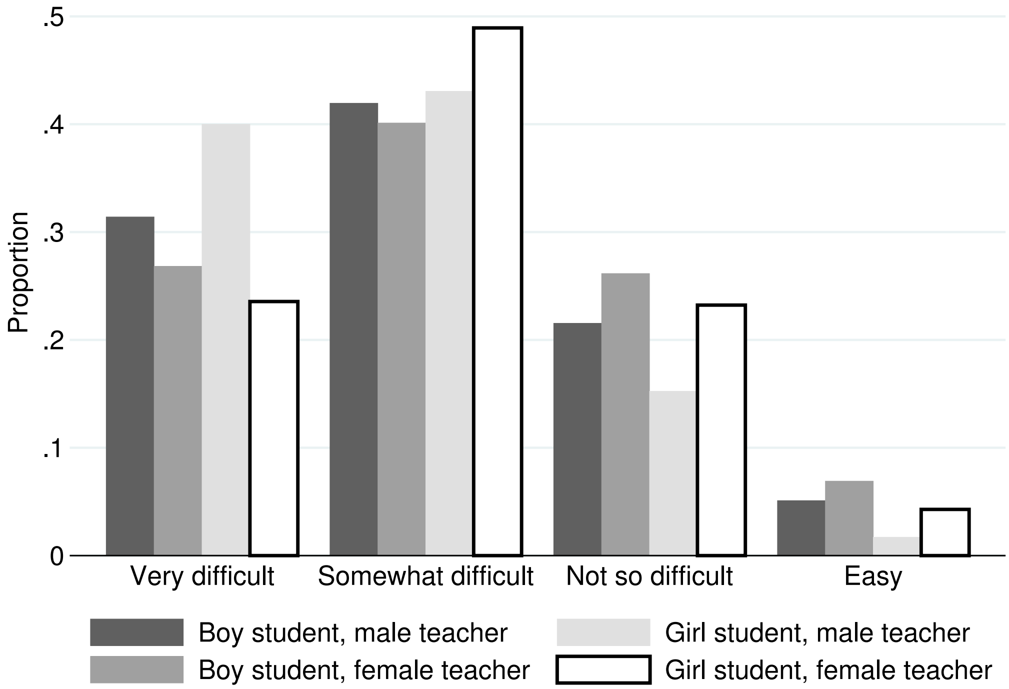
Panel A: Low perceived ability



Panel B: Not low perceived ability

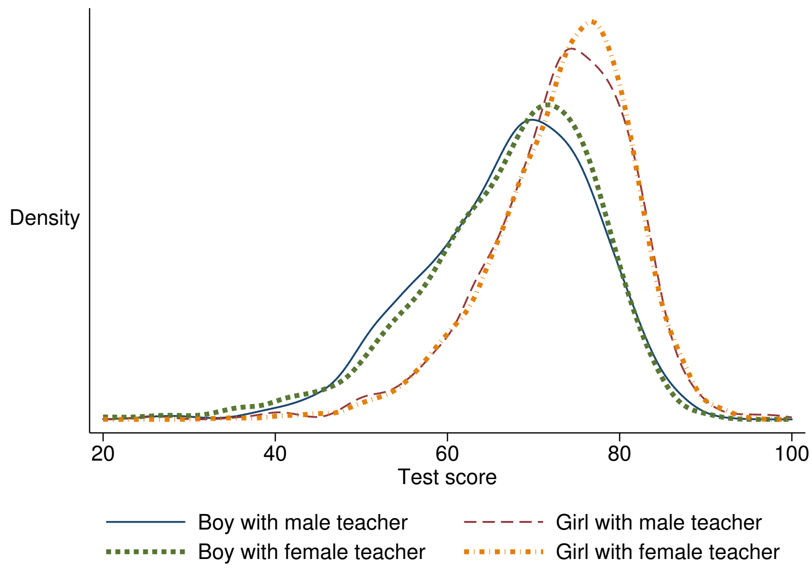
Notes: to generate these figures, we regress test scores on the vector of student-level predetermined characteristics and, using these coefficients, generate a predicted test score for each student. We then plot these using a gaussian kernel for each perceived ability-gender group.

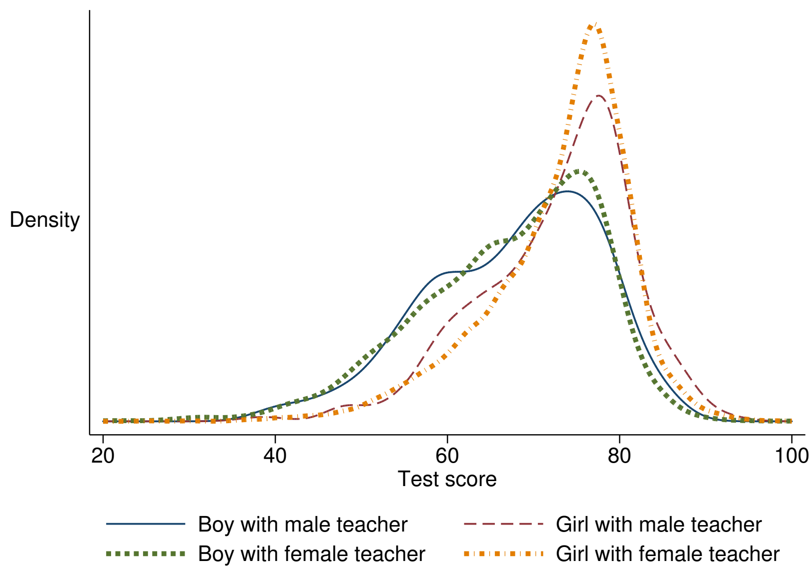
Figure A.3: Effect of teacher-student gender match on student beliefs, for those below within-group median test score



Notes: this figure shows the same analysis as reported in Figure 3, only limiting the sample instead to those below the within-group median math test score.

Figure A.4: Distribution of English and Chinese test scores by teacher-student gender pairing





Notes: this figure shows the analogue to Figure 2 for English and Chinese scores. Note that the large difference between girls assigned to female teachers and all other pairings does not appear hear to the same extent that it does for math scores.

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. (Forthcoming), also exploit this quasi-random assignment of students to classes in Chinese middle schools.