Journal of Population Economics (2025) 38:45
https://doi.org/10.1007/s00148-025-01102-6

ORIGINAL PAPER

On the origins of gender gaps in education: stereotype as a self-fulfilling prophecy

Mingxing Huang[1,2] · **Hongmei Yi**[1,2] · **Scott Rozelle**[3]

Received: 17 February 2024 / Accepted: 14 April 2025 / Published online: 6 May 2025

Abstract

Using reading performance data from a randomized controlled trial of 5224 fifth-grade students in East China, this paper provides a novel test of the hypothesis that evoking a gender stereotype creates gender gaps in education through self-fulfilling prophecies. We found that without intervention, boys performed worse than girls did in reading tests. Evoking a gender stereotype by indicating the expected outperformance of girls over boys in reading had a significantly negative effect on boys and an insignificant effect on girls. As a result, the net effect on the gender gap in reading performance was economically important but statistically insignificant. We also found evidence that increased anxiety was likely the underlying mechanism. Finally, a heterogeneous analysis showed that boys from environments with biased gender role beliefs were more susceptible to the intervention.

Keywords Gender gaps in education · Gender stereotypes · Reading performance · Randomized controlled trial

JEL Classification I21 · I24 · J16

Responsible editor: Alfonso Flores-Lagunes

✉ Hongmei Yi
hmyi.ccap@pku.edu.cn

Mingxing Huang
mx.huang@stu.pku.edu.cn

Scott Rozelle
rozelle@stanford.edu

[1] School of Advanced Agricultural Sciences, Peking University, Beijing 100871, China

[2] China Center for Agricultural Policy, Peking University, Beijing 100871, China

[3] Center On China's Economy and Institutions, Stanford University, Stanford, CA 94305, USA

1 Introduction

It is widely acknowledged that gender gaps exist in early educational achievement. For example, international large-scale assessments have shown that boys score higher on mathematics tests and girls score higher on reading tests (Machin and Pekkarinen 2008; Mullis et al. 2017; OECD 2019). Gender gaps in education in the early stages of one's life can have long-term effects on the educational attainment of children and, ultimately, on their labor market outcomes. Girls that underperform in early math are less likely to pursue advanced math and science courses in high school, leading to their underrepresentation in STEM degrees and STEM occupations (Lavy and Sand 2018; Del Carpio and Guadalupe 2022). Similarly, the gender gap in early reading and spelling skills can be a powerful predictor of the gap in school achievement and the choice of secondary education between boys and girls (Savolainen et al. 2008). Boys that lag behind girls in early reading are at a significantly higher risk of lower educational attainment and ultimately in their levels of income in adulthood than those without such gaps, with childhood reading disabilities reducing the likelihood of achieving higher levels of education and income levels later in life by 74% and 56%, respectively (McLaughlin et al. 2014).

While biological differences between genders may serve as a natural explanation, many studies have investigated the impact of environmental factors, including family, school, and cultural characteristics, on gender gaps in education (Cobb-Clark and Moschion 2017). This study aims to add new evidence to the literature by estimating the causal effect of evoking gender stereotypes in schools on gender gaps in education.[1] Specifically, we designed and conducted a randomized controlled trial of 5224 fifth-grade students in 121 classes from 45 primary schools in a province in East China. A 30-min standardized reading test was administered to each of the students to measure their reading performance. The test was constructed by professional psychometricians based on items from the Progress in International Reading Literacy Study (PIRLS) and has been used in rural China (Gao et al. 2018; Yi et al. 2019). We explicitly evoke gender stereotypes by indicating the expected outperformance of girls over boys in reading before the standardized test. The intervention was randomly assigned to half of the students in the classroom. Relying on a randomized controlled trial to analyze this issue enables us to isolate the role of gender stereotypes from all other factors.

Our main findings are as follows. First, in the control group, where students were not exposed to the intervention (baseline case), boys performed worse than girls did in the reading test, which is consistent with the literature. Then, when analyzing the data from the randomized controlled trial, we found that evoking gender stereotypes by telling students just before the test that girls should be expected to outperform

[1] The Oxford English Dictionary defines a stereotype as a "widely held but fixed and oversimplified image or idea of a particular type of person or thing." The economic approach sees stereotypes as a manifestation of statistical discrimination: rational formation of beliefs about a group member in terms of the aggregate distribution of group traits (Phelps 1972; Arrow 1973). For example, a widely held gender stereotype in education is that "girls are bad at math" (Bordalo et al. 2016).

boys in reading actually reduced the total scores of boys but did not affect girls. However, despite its economic importance, the treatment effect of evoking gender stereotypes on the gender gap is not statistically significant. The significant reduction in the reading performance of boys, without any impact on girls, suggests that the simplicity of the intervention does not undermine its relevance. More importantly, unlike our one-time intervention, real-life stereotype nudging often occurs implicitly and repeatedly throughout the development of children, driven by families, teachers, and societal norms—even without full awareness. Such ongoing reinforcement could have a deeply concerning and long-lasting effect.

We also investigate the mechanisms driving the impact of the intervention. We found that evoking a gender stereotype favoring girls significantly diminished the reading performance of boys, mainly due to increased anxiety rather than reduced confidence. Specifically, although the intervention had negligible effects on the reading confidence of boys, it significantly impaired their performance on more challenging reading tasks, suggesting that heightened anxiety and cognitive load played crucial roles.

Furthermore, this study provides suggestive evidence on how the effect of evoking a gender stereotype varies across students with different families, schools, and cultural backgrounds. We collected detailed information on environmental factors from each student's primary caregiver, the Chinese language teacher, and the school principal. Our findings indicate that the negative effect of gender stereotypes was more pronounced among boys from low-SES and boy-biased families. Similarly, the intervention had a significantly larger negative impact on the reading performance of boys with a senior Chinese language teacher. However, we did not observe significant variations in effect based on family SES, teacher gender, or hukou type. In sum, these results suggest that students from environments with biased gender role beliefs are more susceptible to gender stereotypes.

Through this study, we make three contributions to the literature. First, our study contributes to the literature on explaining gender differences in education. It is fair to say that the literature has done better at documenting the existence of a gender gap in education than at explaining its source (see Cobb-Clark and Moschion (2017) for helpful reviews). Our randomized controlled trial allows us to detect the causal effect of gender stereotypes on gender differences in early educational achievement without interference from other confounding factors that could also generate this gap. While we are not the first to suggest that gender stereotypes play a role in creating gender gaps (for example, Nosek et al. (2009) demonstrated a correlation between national gender stereotypes and educational disparities), the challenge of reverse causation has made it difficult to establish a clear causal link. Our research design overcomes this challenge, providing a robust examination of the causal impact of gender stereotypes on educational outcomes.

Second, this study addresses research on stereotype threats. Stereotypes can lead to self-fulfilling prophecies—a process through which an original expectation leads to its own confirmation, which, in the literature, is called a stereotype threat (Spencer et al. 2016). Existing studies have investigated stereotype threats held by external environments such as school teachers (Alan et al. 2018; Carlana 2019), but few studies have examined the effect of evoking a gender stereotype directly among

students on their performance. In addition, while much of the existing research highlights how girls are affected by stereotype threats in mathematics (Spencer et al. 1999; Lippmann and Senik 2018), there is comparatively less focus on how boys are impacted by language-based stereotype threats. Although the literature is more limited in this area, recent studies in social psychology, such asPansu et al. (2016), Muntoni and Retelsdorf (2018), Li and McLellan (2021), and Chaffee et al. (2024), have begun to explore this issue. Our study not only provides new evidence supporting the self-fulfilling nature of stereotypes but also expands the conversation by examining stereotype threats where gender stereotypes favor girls. This complements existing research on stereotype threats where gender stereotypes favor boys, offering a more balanced view of how these dynamics affect both genders.

Finally, we provide supportive evidence for the hypothesis that variation in gender disparities across families, schools, and cultures may be due to gender stereotypes shaped by different contexts. Although some studies have proposed this hypothesis, it has not yet been tested, possibly because of data limitations (Penner and Paret 2008; Cobb-Clark and Moschion 2017).

The rest of the paper proceeds as follows: Sect. 2 presents a review of the literature. Section 3 introduces the research design, data collection, and empirical strategy. Section 4 reports the main results. Section 5 examines the heterogeneous effects in our effort to explore possible sources of gender stereotypes. We conclude the paper in Sect. 6.

2 Literature review

In this section, we first review the literature on how to evoke a gender stereotype, which our research design mainly refers to. Then, we summarize the existing findings in the literature on how gender stereotypes might vary across family, school, and cultural characteristics. This will guide our heterogeneity analysis in Sect. 5.

2.1 Evoking a gender stereotype in experimental environments

In the social psychology literature, a stereotype threat is often described as a "threat in the air," an invisible force that those affected may not consciously recognize in their daily lives (Steele 1997). A gender stereotype threat, for example, can be triggered by any situational cue that reminds individuals that they are being judged in light of commonly held gender role beliefs. Since people are susceptible to cues that devalue their group identity, situational cues do not necessarily need to be blatant to have an impact (Spencer et al. 2016). In essence, stereotype threats often are subtle, everyday challenges that can affect students without their full awareness.

To evoke gender stereotypes and examine the existence of a stereotype threat in experimental settings, social psychologists commonly have used two methods: informing participants that a particular test has historically shown gender gaps or presenting the test as a diagnostic tool for assessing abilities. For example, to explicitly evoke gender stereotypes, Spencer et al. (1999) told students (before the test

was taken) that gender gaps were expected to emerge from the math examination. Steele and Aronson (1995) and Pansu et al. (2016) used more subtle cues, describing the test as a measure of academic ability to implicitly create a stereotype-threat condition.

In this study, we adopted the explicit method used by Spencer et al. (1999). Specifically, we first showed the gender gaps of students that had been identified in past reading tests. Immediately afterward, we asked the students to confirm their own gender. This step is based on findings by Danaher and Crandall (2008), which showed that simply indicating one's gender before a test can evoke gender stereotypes and trigger a stereotype threat.

2.2 Internalized gender role beliefs and heterogeneous effects of evoking gender stereotypes

Evoking a gender stereotype is more likely to harm children who have grown up in environments with more explicitly biased gender role beliefs. Children in such environments typically internalize biased gender role beliefs through long-term exposure to gender stereotypes (Steele 1997). As a result, gender stereotypes, which are evoked by situational cues, confirm the internalized beliefs of those children that boys and girls are different due to their gender roles, which in turn increases their likelihood of following what the gender stereotype has prescribed and becoming more vulnerable to stereotype threats (Aronson et al. 1999). In the process of child development, families, teachers, and cultural environments all play critical roles in shaping and reinforcing the beliefs of children about gender roles, thus becoming possible sources of gender stereotypes.

Families can have the earliest and most direct impact on the gender role beliefs of children in their early stages of life. Any gender-specific behavior of parents induced by son preference (for example) may embody implicitly biased gender role beliefs, which could affect the way their children think and behave.[2] For example, first-born girls with younger brothers are more likely to internalize traditional gender roles due to differential parenting (Brenøe, 2021). Dossi et al. (2021) also found that girls from boy-biased families—where all children are girls except for the last-born child—scored significantly lower on math tests than did those raised in non-boy-biased families. Additionally, son preference has also been shown to be related to the unequal distribution of scarce educational resources between boys and girls within the family. In credit-constrained or low-socioeconomic-status (low-SES) families, boys are often given more educational resources than girls (Rose 2000; Lei et al. 2017). In contrast, in a high-SES family, parents are able to provide abundant education resources to all their children; thus, there is less gender stereotyping (Lily 1994).

Teachers also significantly influence student gender role beliefs. Previous studies have tested and verified the presence of gender stereotypes in schools (Lavy

[2] Previous studies have shown that son preference is widespread in the world including but not limited to China, India, and the US (Das Gupta et al. 2003; Dahl and Moretti 2008; Rosenblum 2013).

2008; Alan et al. 2018; Carlana 2019). Because teachers are authority figures, students tend to follow what the teacher says and internalize the gender role beliefs that their teachers hold (Alan et al. 2018). Two significant factors contributing to this are the teacher gender and their years of teaching. Regarding the teacher gender, prior research suggests that gender matches between teachers and students influence educational outcomes in the capacity of a gender role model (Bettinger and Long 2005; Dee 2007; Carrell et al. 2010; Muralidharan and Sheth 2016). However, having a role model could produce a positive effect only if the teacher is successful and performs well (Beilock et al. 2010; Antecol et al. 2015). In terms of teaching experience, studies indicate that teachers with more seniority tend to have stronger gender stereotypes because they often make judgments about current students based on previous teaching experience. This process implies a form of confirmation bias in light of new information: beliefs are formed that overreact to information that confirms the stereotype and ignore information that contradicts it (Bordalo et al. 2016).

Finally, culture, which is defined as beliefs and values passed down through generations, directly shapes the gender role beliefs of children (Guiso et al. 2006). These gender role beliefs, rooted in the past, continue to influence educational gender gaps. For example, in Italy, the commercial culture of the late Middle Ages helped narrow the gender gap in education by valuing women's education (Bertocchi and Bozzano 2016). Similarly, China's long-standing agricultural culture, which favored men due to the physical demands of plow agriculture, has perpetuated unequal gender norms (Alesina et al. 2013). While brawn-intensive work practices are still favored in rural areas, many traditional Chinese agricultural practices have disappeared in recent years and are no longer present in urban areas. Thus, gender stereotypes are more likely to exist in rural areas than in urban areas (Istenič 2007).

3 Methodology

3.1 Sampling

We designed and conducted a randomized controlled trial (RCT) of fifth-grade students from 121 classes in 45 primary schools in Anhui Province. We chose fifth graders who were mostly 10 and 11 years old since the age range of 9 to 13 years represents a crucial period for reading development (Fitzgerald and Shanahan 2000), and early-emerging gender gaps in education can be persistent. In 2020, the per capita GDP of Anhui Province was 63,426 yuan (approximately 9195 USD), which was 13 th among the 31 mainland provinces (CNBS 2021). Primary school students in Anhui Province account for 4.4% of all primary students in China (CNBS 2021).

Our data consist of two separate subsamples. The first subsample includes all 36 primary schools in county A of Hefei city (urban schools), the capital of Anhui Province. The sampling procedure involved three steps. First, we collected a detailed list that included class names and the number of students in each class in the fifth grade from all primary schools from the local education department of County A. In total, there were 9692 students in 217 classes in the fifth grade from 36 schools in September 2020. Second, we randomly selected three classes from each school.

If there were fewer than three classes in a grade, we selected all the classes. Third, we surveyed all the students in the selected classes. The survey and RCT were conducted in December 2020 in these urban schools.

The second subsample includes nine rural primary schools in Anhui Province. The selection process of rural schools differed from that of urban schools. First, using data from urban surveys, we identified the source counties from which migrant students in urban schools originated. Subsequently, we selected the three top counties according to the frequency of the migrant students. Then, in each source county, we divided the townships into three groups according to economic development (by tercile) and randomly selected one township from each of the three tercile groups. Finally, we selected the central primary school in each township as part of our school sample. The sampling procedure within each of the selected rural schools was the same as that used for the urban schools. The same survey and RCT were administered to the selected students in the rural schools in March 2021, which ensured that our interventions in the two subsamples were essentially consistent.[3]

The sampling procedure resulted in a sample of 5331 students in 121 fifth-grade classes from 45 schools. Of these, 101 classes and 4456 students were from the 36 urban schools, and 20 classes and 875 students were from the nine rural schools.

3.2 Intervention and randomization

The objective of the intervention was to evoke a gender stereotype among students by instructing them that a particular test had shown gender gaps in the past. Specifically, we administered a standardized test to students in the classroom where the students were told that the tests were a diagnosis of academic ability. We designed two versions of the reading test, which we named the type-A test and the type-B test. The only difference between the type-A and type-B tests was that we included an additional statement and a question at the top of the type-B test. The statement was as follows: *Studies have shown that primary school girls score higher on reading tests than boys in most countries*. This statement was included to show students that there are gender gaps that have been identified based on past reading tests. Immediately afterward, we asked students to confirm their own gender by asking, "*What is your gender? (A. Girl; B. Boy).*"

In each classroom, students were randomly assigned to take either the type-A test or the type-B test. Two enumerators administered the tests inside the classroom with the assistance of the homeroom teacher from each sample class. As soon as the survey team entered the classroom, the enumerators confirmed the number of students who would be participating in the tests. If the number of participants was even, half of the students were given the type-A test, and the other half were given the type-B test at random. If the number was odd, the extra student was randomly assigned one of the two types of tests. All the students were required to finish the reading test within half an hour. The tests were also closely monitored by the survey team, and students had no chance to read (or copy) the tests of other students in the class. All the students were blind to the random assignment of the intervention.

[3] The time gap between the urban and rural surveys occurred because we received additional funding to conduct the rural survey after completing the urban survey.

We acknowledge that the RCT was conducted during the COVID- 19 pandemic, a factor that could potentially influence our findings. However, it is important to note that our intervention was administered within individual classrooms, ensuring that COVID- 19 had a uniform impact on both treatment and control groups. This design minimizes concerns about differential effects related to the pandemic, including the timing of school closures, as both groups were subject to the same conditions.

3.3 Data collection

We used the same survey instruments to collect data across two subsamples. Surveys were administered to students, their primary caregivers, their Chinese language teachers, and the school principals. The survey instruments were then sent to the local education department for review and approval before the field survey.

The student survey consisted of two parts. First, a 30-min standardized reading test was administered to each of the students to measure their reading performance. The test, which has been used in rural China, was constructed by professional psychometricians based on items from the Progress in International Reading Literacy Study (PIRLS) (Gao et al. 2018; Yi et al. 2019). The items could be classified into four groups based on the *comprehension processes* of reading: the ability to focus on and retrieve explicitly stated information; the ability to make straightforward inferences; the ability to interpret and integrate ideas and information; and the ability to examine and evaluate content, language, and textual elements (Mullis et al. 2012). In the second part of the student survey, we asked each student in the selected classes to complete an e-questionnaire under the guidance of the two enumerators. The questionnaire collected data on student demographic characteristics such as gender, age, and academic performance during the previous semester. In the last part of the student survey, we assessed the affinity of each student for reading, and student confidence in reading using two scales from the PIRLS surveys (Martin et al. 2017).

We also administered an e-questionnaire to the primary caregiver of each student. The primary caregiver was defined as the adult person who had the main responsibility for caring for each student's studies and daily life. The primary caregiver survey collected information on the characteristics of the family of each student, including hukou registration,[4] the highest education level of the sample student's family members, the family's annual per capita income, and information about the siblings of the student. Detailed information on each student's siblings included the number of his or her older brothers/older sisters/younger brothers/younger sisters.[5] This question allowed us to produce a measure of each family's preference for

[4] Hukou is a household registration system that is designed to control population movement in China and is closely connected to the rights that an individual has to benefit from public services. The two types of hukou, namely, urban and rural, pertain to urban and rural population, respectively (Liu 2005). In recent years, China has undertaken hukou reforms to relax internal migration restrictions (Zhang et al. 2024).

[5] It is very likely that many primary caregivers in rural areas could not fill in the survey form due to working away from home or illiteracy in writing. Thus, we collected information about the siblings of the sample student through the student survey (instead of primary caregiver survey) in the rural survey.

boys—*Son preference* (categorized as non-boy-biased families; boy-biased families; and families with indeterminate preferences for boys).

Finally, we collected information on the gender and age of the Chinese language teachers through the use of surveys. The principal questionnaire collected information on school-level characteristics.[6]

Of the 5331 students selected, 5224 completed the survey in December 2020 or March 2021. The main reasons that students failed to complete the survey were absence on the survey day or inability to complete the survey independently due to a disability. Of the students who completed the survey, 4367 were from 101 classes in the 36 urban schools, and 857 were from 20 classes in the nine rural schools.

The descriptive statistics indicate that randomization leads to a balanced sample. Specifically, by our randomization, 2605 students were assigned to the treatment group, and 2619 students were assigned to the control group. Approximately 54% of the students in each group were boys. According to our within-classroom design, there was no statistically significant difference between the treatment group and the control group across a wide range of student and family characteristics (Table 1).

We further examined whether the missing data biased our analysis. Specifically, of the 5224 students who completed the student questionnaires, only 4722 (90%) completed the caregiver questionnaires. Thus, when we conducted the analysis using variables from this survey (i.e., family education, family income, and the number of siblings), the results might be biased in terms of the representativeness of the sample or the nonrandom attrition. To address this problem, we conducted three types of analysis. First, we tested whether the missing family data were related to random assignment and found that students in the treatment group were more likely to have missing values than were those in the control group by two percentage points. However, given the high completion rate (90% versus 91%), the magnitude of the difference was negligible (Panel A1, Appendix Table A1). Furthermore, we tested whether the student characteristics in the treatment group were significantly different from those in the control group based on non-missing data. The results indicated that after excluding observations with missing data, the two groups were statistically indifferent (Panel A2, Appendix Table A1). Finally, we tested whether the students with missing primary caregiver data were significantly different from those with complete information. The results suggested that students with non-missing data are more likely to represent high-performing students (Panel A, Appendix Table A2). The analysis on the missing of the number of siblings shows similar results (Panels B1 & B2, Appendix Table A1; Panel B, Appendix Table A2). In general, although this is the case, the potential bias due to missing data will be limited.

[6] Program information and full texts of questionnaires are available via the website: http://scholar.pku.edu.cn/hongmei-yi/program/reading.

Table 1 Balancing test

	Control group		Treatment group		Difference	
	Mean	Std. dev	Mean	Std. dev	Difference	*p* value
	(1)	(2)	(3)	(4)	(5)	(6)
Student characteristics:						
Male (yes = 1)	0.54	0.50	0.54	0.50	− 0.00	0.94
Age (year)	11.36	0.55	11.36	0.56	0.00	0.87
Chinese language test performance last semester (A = 1)	0.40	0.49	0.40	0.49	0.00	0.82
Math test performance last semester (A = 1)	0.58	0.49	0.60	0.49	0.02	0.23
Observations	2619		2605			
Family characteristics:						
Student received preschool education (yes = 1)	0.65	0.48	0.63	0.48	− 0.02	0.19
Rural hukou (yes = 1)	0.41	0.49	0.41	0.49	− 0.01	0.61
High-education families (highest education level of family members is above high school) (yes = 1)	0.48	0.50	0.47	0.50	− 0.01	0.67
High-income families (annual income per capita > 100,000 yuan) (yes = 1)	0.30	0.46	0.28	0.45	− 0.01	0.25
Observations	2390		2332			
Characteristics derived from each student's sibling information:						
Birth order of the student						
First-born child (1 = yes)	0.68	0.47	0.69	0.46	0.01	0.39
Second-born child (1 = yes)	0.26	0.44	0.25	0.43	− 0.01	0.45
Third-born child (1 = yes)	0.07	0.25	0.07	0.25	− 0.00	0.85

Table 1 (continued)

	Control group		Treatment group		Difference	
	Mean	Std. dev	Mean	Std. dev	Difference	*p* value
	(1)	(2)	(3)	(4)	(5)	(6)
Son preference of the family						
Non-boy-biased families (yes = 1)	0.43	0.49	0.41	0.49	− 0.01	0.33
Boy-biased families (yes = 1)	0.24	0.43	0.26	0.44	0.01	0.24
Families with unidentifiable preferences for boys (yes = 1)	0.33	0.47	0.33	0.47	0.00	0.93
Observations	2442		2386			

Notes: Student characteristics were collected from the student e-questionnaire, family characteristics were collected from the primary caregiver e-questionnaire, and characteristics derived from each student's sibling information were collected from the primary caregiver e-questionnaire in the urban survey and the student e-questionnaire in the rural survey. Columns (1) and (3) present the mean of each variable, and columns (2) and (4) present the standard deviation. Column (5) presents the difference, and column (6) presents the *p* value with standard errors clustered at the class level. Specifically, in columns (5) and (6), each row represents a separate regression, in which the independent variable is an indicator for the treatment group, and the dependent variable is the corresponding student, family and sibling characteristics as listed above. All specifications in columns (5) and (6) include class fixed effects

Inference: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

3.4 Estimation strategy

3.4.1 Main analysis

We used the double-difference regression model to estimate the treatment effect of evoking a gender stereotype on student reading performance (Model 1):

$$Y_{ic} = \alpha + \beta_1 Male_{ic} + \beta_2 Treat_{ic} + \beta_3 \left(Male_{ic} \times Treat_{ic}\right) + I'_{ic}\gamma + F_{ic}\prime\delta + D_c + \varepsilon_{ic} \quad (1)$$

where Y_{ic} is the reading test score of student i in class c. In this study, the overall total scores and the scores for four reading comprehension processes were all measured and normalized to z scores, with a mean of zero and a standard deviation of one. The descriptive statistics of these outcome variables are reported in Appendix Table A3. $Male_{ic}$ and $Treat_{ic}$ indicate the gender and treatment status of student i, respectively. $Male_{ic}$ takes the value of one if the student is male, and $Treat_{ic}$ equals one if the student is from the treatment group. Otherwise, the variable takes the value of zero. D_c is a series of class dummies to account for our within-classroom design. In order to improve the efficiency of estimation, we also include controls for student characteristics (I_{ics}) and family characteristics (F_{ics}) in the model. Controls of family characteristics include the number of siblings, birth order, high-education families (if the highest education level of family members is above high school, it equals 1; otherwise, 0), high-income families (if annual income per capita is more than 100,000 yuan, it equals 1; otherwise, 0), rural hukou (yes = 1), and whether the student received preschool education (yes = 1). Controls of student characteristics include the Chinese language test performance last semester ($A = 1$) and the math test performance last semester ($A = 1$). ε_{ic} denotes the error term. In all regressions in this study, we accounted for the clustered nature of our sample by correcting standard errors for class-level clustering.

β s are the coefficients of interest. First, β_1 represents the gender gap in reading achievement in the context without the intervention of gender stereotypes.[7] Second, β_2 and $\beta_2 + \beta_3$ represent the treatment effects of gender stereotypes on the reading performance of female students and male students, respectively. Finally, β_3 captures the treatment effects of gender stereotypes on the gender gap in reading performance.

3.4.2 Heterogeneity analysis

Our heterogeneous analysis is guided by the relevant literature discussed in Sect. 2.2. Specifically, we estimate the heterogeneous effects of evoking a gender stereotype on students from different families, from different classes, and from different cultural backgrounds. We conduct subsample regressions based on heterogeneous variables. To formally test the differential impacts, we perform a seemingly unrelated

[7] The gender gap is calculated by scores of boys minus that of girls, and henceforth.

test and report the p values for the regression coefficients. As a robustness check, we also use a triple-difference regression model and report the results in the Appendix.

We first try to examine the heterogeneous effect of the son preference of the family. To contextualize the definition of a "boy-biased" family, it is essential to consider China's fertility policies, particularly the relaxation of the family planning policy. In response to a number of demographic challenges, China began gradually relaxing the One-Child Policy starting in 2000, allowing families where both parents were only children to have a second child. In 2013, the policy was relaxed nationwide to allow couples where either of the couple was an only child to have two children. Policymakers then implemented an overall Two-Child Policy in 2016, allowing all families to have two children. More recently, this policy was relaxed even further—to a Three-Child Policy.[8] In light of these policy changes, families have gained greater flexibility in determining the number and gender composition of their children. In our sample, only 32% of students are the only child in the family. We validated the son preference in China by observing that fertility is higher for those families where the first born is a girl (see Appendix Table A4 for results, and the process to identify the gender of the first-born child is reported in Appendix Table A5). Then, following the spirit of Dossi et al. (2021), we built a measure of son preference based on family fertility patterns (Appendix Table A6). Specifically, boy-biased families are those where all children are girls except for the last born. Families with unidentifiable preference for boys are those where there is only one child (and it is a boy), and those where the surveyed student is a girl and she has both younger brothers and sisters (although we cannot identify the gender of the last-born child). Non-boy-biased families are all the other types of families. This definition, while more applicable in contexts without fertility restrictions, remains relevant in our sample due to the recent relaxation of China's family planning policies. In our sample, 1203 families (25%) were identified as having boy-biased preferences, 2027 families (42%) were identified as not having boy-biased preferences, and 1598 students (33%) were from families with unidentifiable son preferences. The families were equally distributed in the treatment and control groups (Table 1).

In the heterogeneous analysis, we run two regressions using model (1) with the subsamples from boy-biased families and other types of families (including non-boy-biased families and families with unidentifiable preference). As a robustness check, we further focused on students whose families have identifiable son preferences (boy-biased families versus non-boy-biased families) and reported results in the Appendix. We also explored whether the heterogeneous effect of son preference on families with low socioeconomic status (i.e., low education and low income) significantly differed from that on families with high socioeconomic status. The distribution of families with different son preferences by family socioeconomic status and treatment status is presented in Appendix Figure A1 and suggests insignificant differences between subgroups ($p > 0.10$).

We then examined two sets of heterogeneous effects at the class level: gender and seniority. The Chinese language teacher is assumed to play the largest role in

[8] See Li and Shi (2025) and Huang et al. (2025) for more details on the fertility policies in China.

Table 2 The effect of evoking a gender stereotype on student reading performance

	Normalized total scores (SD)			
	(1)	(2)	(3)	(4)
Male (β_1)	− 0.064 (0.043)	− 0.084** (0.040)	− 0.084** (0.042)	− 0.046 (0.041)
Treatment (β_2)	0.003 (0.047)	− 0.013 (0.039)	− 0.004 (0.040)	− 0.007 (0.040)
Male × treatment (β_3)	− 0.086 (0.067)	− 0.060 (0.055)	− 0.081 (0.057)	− 0.088 (0.057)
Constant (α)	0.056 (0.050)	− 0.011 (0.024)	− 0.073* (0.039)	− 0.422*** (0.040)
Observations	5224	5224	4722	4722
R-squared	0.004	0.253	0.263	0.312
Class fixed effects	No	Yes	Yes	Yes
Controls for family characteristics	No	No	Yes	Yes
Controls for student characteristics	No	No	No	Yes
F-test of coefficients of interest				
$\beta_2 + \beta_3$	− 0.083** (0.041)	− 0.073* (0.039)	− 0.085** (0.038)	− 0.096** (0.037)

Notes: Dependent variables are normalized total scores (SD). Controls for family characteristics include the number of siblings, birth order, high-education families (highest education level of family members is above high school) (yes = 1), high-income families (annual income per capita > 100,000 yuan) (yes = 1), rural hukou (yes = 1), and student received preschool education (yes = 1). Controls for student characteristics include Chinese language test performance last semester ($A = 1$) and math test performance last semester ($A = 1$). Standard errors are clustered at the class level and reported in parentheses

Inference: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

developing the reading skills of the sample students. We used 40 years of age as a cutoff to define whether a teacher was senior. Of the 121 classes in our sample, 16 (13%) had a male Chinese language teacher, and 51 (42%) had a senior Chinese language teacher.

Finally, we sought to measure the nature of the heterogeneity at the cultural level—the rural culture. We used students' hukou status (*rural* or *urban*) as a proxy of rural culture. In our sample, 41% of caregivers reported that the student was registered as a rural hukou and was equally distributed between the treatment and control groups (Table 1).

4 Main results

4.1 Gender gap and stereotype impact on student reading performance

Before examining the impact of evoking a gender stereotype, we first compared the gender differences in reading performance in the control group (β_1 in Table 2).[9]

[9] The raw differences in mean (without controlling for class fixed effects) between boys and girls in the control group are presented in Appendix Figure A2 and Appendix Figure A3.

Without any inclusion of controls, boys scored 0.064 SD lower than girls, and this difference was not statistically significant. However, when we controlled for class fixed effects, the gender gap widened to 0.084 SD and became significant at the 5% level. After accounting for family characteristics, this gap remained largely unchanged. These results suggest a significant gender gap in reading performance prior to our intervention, even among students in the same classroom with similar family backgrounds. This finding is consistent with the results from the 2016 PIRLS, where researchers found that boys perform significantly and substantially worse than girls (Mullis et al. 2017). Nevertheless, it is not surprising that the gender difference disappeared when we further controlled for student previous academic performance.

Next, we investigated the effect of evoking a gender stereotype on the reading performance of boys and girls, examining its impact on the gender gap in reading.[10] Our findings revealed a significant and robust negative stereotype effect on the reading performance of boys ($\beta_2 + \beta_3$ in Table 2). In the absence of controls, boys in the treatment group—where boys are the stereotyped gender in our intervention—performed significantly worse than those in the control group (− 0.083 SD, $p < 0.05$). This negative effect persisted after controlling for class fixed effects and family characteristics, with the effect size at −0.073 SD ($p < 0.10$) and −0.085 SD ($p < 0.05$), respectively. Notably, after accounting for student prior academic performance, the negative impact of the stereotype intervention even increased to −0.096 SD ($p < 0.05$). These results suggest that the stereotype intervention significantly diminishes the reading performance of boys.

The intervention had a negligible and statistically insignificant effect on the reading performance of girls (β_2 in Table 2). Without any inclusion of controls, girls in the treatment group performed nearly the same as those in the control group (0.003 SD, $p > 0.10$). After accounting for class fixed effects, family characteristics, and student characteristics, the effect remained negligible, with coefficients of −0.013 SD ($p > 0.10$), −0.004 SD ($p > 0.10$), and − 0.007 SD ($p > 0.10$), respectively. The impact of the stereotype intervention on girls was insignificant both statistically and economically, in contrast to its effect on boys.

Overall, the intervention widened the gender gap in reading, although the effect was statistically insignificant (β_3 in Table 2). Initially, evoking a gender stereotype enlarged the gender gap by −0.086 SD. We obtained similar results after adjusting for class fixed effects, family characteristics, and student characteristics, with coefficients of −0.060 SD, −0.081 SD, and −0.088 SD, respectively. Although the stereotype effect on the gender gap approached statistical significance (e.g., $p = 0.123$ for β_3 in column 4 of Table 2), the effect size was moderate (or slightly below average) compared to typical findings in stereotype threat research, which generally report average effect sizes exceeding 0.20 SD (Nguyen and Ryan 2008; Flore and Wicherts 2015).

We also compare our effect size to those reported in the broader education literature to better situate our findings. Systematic reviews indicate that educational

[10] The raw differences in mean (without controlling for class fixed effects) between boys and girls in the control group are presented in Appendix Figure A2 and Appendix Figure A3.
nd Appendix Figure A5.

interventions typically yield a median impact of 0.10 SD on achievement outcomes like reading (Evans and Yuan 2022; Kraft 2023), and our effect size aligns closely with this benchmark. We also compare our effect size with that of Yi et al. (2019) and Gao et al. (2022), two papers that used the same reading tests as our study. Yi et al. (2019) found that their in-class library intervention had an effect size of less than 0.01 SD, while Gao et al. (2022) reported effect sizes of −0.09 SD, 0.09 SD, and 0.61 SD for three interventions: providing a book corner, providing a book corner with training from the education bureau, and providing a book corner with NGO-led training, respectively. Given the substantial costs of these reading programs, the fact that our simple intervention—adding just one line before the reading test—can yield such a significant impact is noteworthy.

Our findings align with previous studies, such as Carlana (2019), which demonstrate that gender stereotypes negatively impact the academic performance of students whose gender is stereotypically viewed as disadvantaged in a particular subject. In our study, this effect was evident in the way gender stereotypes impaired the reading performance of boys, a group often stereotyped as being weaker in reading. Conversely, the performance of girls, who are stereotypically viewed as being stronger in reading, remained largely unaffected by the stereotype intervention. This suggests that while stereotypes can be detrimental to students who are perceived as less capable in certain subjects, they do not necessarily boost the performance of those viewed as more capable. This reinforces the idea that gender stereotypes create gender gaps in education, where the burden of negative expectations falls disproportionately on those already at a disadvantage.

4.2 Mechanisms of the stereotype threat

Our analysis demonstrated that evoking a gender stereotype which favors girls significantly reduced the reading performance of boys. In this section, we explore two potential mechanisms from the literature—lower confidence and increased anxiety—that may explain how this stereotype contributes to boys' underperformance in reading.

4.2.1 Lower confidence

One potential mechanism is that evoking a gender stereotype may lower the confidence of the stereotyped students in their ability to succeed in the subject (Carlana 2019). To examine whether our stereotype intervention discouraged boys in reading, we analyzed data that were collected after the intervention. This included assessments of each student's affinity for reading and their confidence in their reading ability, based on two scales from the PIRLS surveys (Martin et al. 2017): the Students Like Reading Scale and the Students Confident in Reading Scale. These scales were measured and normalized to z-scores, with a mean of zero and a standard deviation of one.

Our analysis shows that, while boys had lower reading affinity and confidence than girls in the control group (β_1 in Table 3), the stereotype intervention did not

Table 3 The effect of evoking a gender stereotype on student reading affinity and confidence

	Normalized Students Like Reading Scale (SD)		Normalized Students Confident in Reading Scale (SD)	
	(1)	(2)	(3)	(4)
Male (β_1)	− 0.274*** (0.043)	− 0.205*** (0.040)	− 0.207*** (0.037)	− 0.147*** (0.038)
Treatment (β_2)	0.014 (0.033)	0.020 (0.030)	0.025 (0.040)	0.044 (0.041)
Male × treatment (β_3)	0.082 (0.058)	0.040 (0.056)	− 0.011 (0.051)	− 0.048 (0.056)
Constant (α)	0.237*** (0.023)	− 0.109** (0.047)	0.084*** (0.024)	− 0.262*** (0.045)
Observations	5224	4722	5224	4722
R-squared	0.156	0.195	0.147	0.209
Class fixed effects	Yes	Yes	Yes	Yes
Controls for family characteristics	No	Yes	No	Yes
Controls for student characteristics	No	Yes	No	Yes
F-test of coefficients of interest				
$\beta_2 + \beta_3$	0.096** (0.045)	0.060 (0.044)	0.013 (0.034)	− 0.004 (0.035)

Notes: Dependent variables are Students Like Reading Scale (SD) and Students Confident in Reading Scale (SD), both of which are measured and normalized to z-scores, with a mean of zero and a standard deviation of one. Higher value indicates a higher level of reading affinity. Controls for family characteristics include the number of siblings, birth order, high-education families (highest education level of family members is above high school) (yes = 1), high-income families (annual income per capita > 100,000 yuan) (yes = 1), rural hukou (yes = 1), and student received preschool education (yes = 1). Controls for student characteristics include Chinese language test performance last semester ($A = 1$) and math test performance last semester ($A = 1$). Standard errors are clustered at the class level and reported in parentheses

Inference: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

further widen this gender gap (β_3 in Table 3). Specifically, for reading affinity (Columns 1 and 2 of Table 3), the coefficients for the stereotype intervention on the reading affinity of boys were positive (0.096 SD and 0.060 SD, respectively, under two specifications). The intervention had a negligible effect on the reading affinity of girls and a positive, though statistically insignificant, effect on the gender gap. For reading confidence (columns 3 and 4 of Table 3), the stereotype intervention showed only negligible effects on the reading confidence of either boys or girls; there also was only a negligible effect on the gender gap. These results suggest that reducing the confidence of students is unlikely to be the underlying mechanism of the stereotype threat observed in our study.

4.2.2 Increased anxiety

In social psychology, a stereotype threat also is thought to occur via increased anxiety and an increased cognitive load created by such anxiety (Steele 1997). When individuals face a stereotype threat, they often strive to disprove negative stereotypes tied to their social identity. This heightened motivation to counter or avoid reinforcing the stereotype adds extra pressure to perform well, which can paradoxically undermine their performance by increasing cognitive load. If this mechanism is at play, we would expect the stereotype intervention to have a more pronounced impact on difficult tests compared to easier ones, since it is during these challenging tasks that the added burden of a stereotype threat would be expected to interfere with their performance (Ben-Zeev et al. 2005).

To examine this, we categorized the reading test into easy and difficult components. According to the PIRLS classification (Mullis et al. 2012), our reading items could be classified into four groups based on the *comprehension processes* of reading: (a) the ability to focus on and retrieve explicitly stated information; (b) the ability to make straightforward inferences; (c) the ability to interpret and integrate ideas and information; and (d) the ability to examine and evaluate content, language, and textual elements. The latter two types of tasks, which require higher-level reading skills, were classified as difficult, while the remaining tasks were classified as easy. As a robustness check, we also classified tasks based on their accuracy rates. Specifically, we calculated the accuracy rate for each item and classified tasks in the bottom 50% of accuracy rates as difficult and those in the top 50% as easy.

Our findings reveal a clear pattern: the stereotype intervention disproportionately harmed boys on the more difficult tasks ($\beta_2 + \beta_3$ in Table 4). Specifically, in the baseline specification (columns 1, 3, 5, and 7 in Table 4), the intervention had no significant effect on boys for the two easy tasks (− 0.051 SD and − 0.047 SD, respectively; $p > 0.10$), but it significantly harmed boys on the two difficult tasks (− 0.083 SD and − 0.092 SD, respectively; $p < 0.05$). This pattern persists even after controlling for family and student characteristics (columns 2, 4, 6, and 8 in Table 4). The intervention reduced the scores of boys on tasks related to interpreting and integrating ideas and information by − 0.098 SD ($p < 0.05$), and on tasks related to examining and evaluating content, language, and textual elements by − 0.114 SD ($p < 0.01$). These effects were significantly larger than those observed for the two easier tasks

Table 4 The effect of evoking a gender stereotype on student reading performance by four comprehension processes of reading

	Normalized scores on focusing on and retrieving explicitly stated information (SD) (easy tasks)		Normalized scores on making straightforward inferences (SD) (easy tasks)		Normalized scores on interpreting and integrating ideas and information (SD) (difficult tasks)		Normalized scores on examining and evaluating content, language, and textual elements (SD) (difficult tasks)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male (β_1)	− 0.091** (0.037)	− 0.059 (0.038)	− 0.114*** (0.042)	− 0.076* (0.041)	− 0.022 (0.044)	0.011 (0.046)	− 0.036 (0.040)	− 0.008 (0.042)
Treatment (β_2)	− 0.009 (0.039)	− 0.010 (0.039)	− 0.013 (0.040)	− 0.007 (0.042)	− 0.014 (0.043)	− 0.000 (0.044)	− 0.010 (0.041)	− 0.004 (0.043)
Male × treatment (β_3)	− 0.042 (0.056)	− 0.060 (0.057)	− 0.034 (0.055)	− 0.065 (0.057)	− 0.069 (0.055)	− 0.098* (0.058)	− 0.081 (0.052)	− 0.110** (0.054)
Constant (α)	0.046** (0.023)	− 0.306*** (0.042)	− 0.031 (0.024)	− 0.462*** (0.042)	− 0.173*** (0.027)	− 0.512*** (0.044)	0.097*** (0.025)	− 0.194*** (0.041)
Observations	5224	4722	5224	4722	5224	4722	5224	4722
R-squared	0.217	0.264	0.212	0.264	0.204	0.249	0.229	0.261
Class fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls for family characteristics	No	Yes	No	Yes	No	Yes	No	Yes
Controls for student characteristics	No	Yes	No	Yes	No	Yes	No	Yes
F-test of coefficients of interest								
$\beta_2 + \beta_3$	− 0.051 (0.040)	− 0.070* (0.039)	− 0.047 (0.039)	− 0.072* (0.037)	− 0.083** (0.040)	− 0.098** (0.042)	− 0.092** (0.041)	− 0.114*** (0.042)

Notes: Dependent variables are normalized scores by four comprehension processes of reading (SD). Controls for family characteristics include the number of siblings, birth order, high-education families (highest education level of family members is above high school) (yes = 1), high-income families (annual income per capita > 100,000 yuan) (yes = 1), rural hukou (yes = 1) and student received preschool education (yes = 1). Controls for student characteristics include Chinese language test performance last semester ($A = 1$) and math test performance last semester ($A = 1$). Standard errors are clustered at the class level and reported in parentheses

Inference: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

(− 0.070 SD and − 0.069 SD, respectively; $p < 0.10$). The pattern remains consistent in the robustness checks ($\beta_2 + \beta_3$ in Appendix Table A7).

Since the intervention had negligible effects on girls across all tasks (β_2 in Table 4) and more negative effects on boys in the difficult tasks, the intervention widened the gender gap more on difficult tasks (β_3 in Table 4). We also note that the initial gender gap was negligible on the difficult tasks compared to the easy ones (β_1 in Table 4). One possible explanation is that both boys and girls performed poorly on the difficult tasks, resulting in no initial gender gap. This pattern also holds in the robustness checks (β_1 in Appendix Table A7).

Overall, these findings suggest that it might be increased anxiety, along with the rising cognitive load that it generates, which may be the underlying mechanism driving the stereotype threat observed in our study.

5 Heterogeneous effects of the intervention

5.1 Heterogeneous effects of the intervention on students from boy-biased and low-SES families

The heterogeneous analysis shows that the intervention had markedly different effects on students from boy-biased families compared to those from other family types (columns 1 to 3 of Table 5).[11] First, in families without boy-biased preferences, the reading performance of boys was significantly lower than the reading performance of girls by − 0.083 SD ($p < 0.10$; β_1 in column 2) in the control group. In contrast, in boy-biased families, the gender gap was positive and economically significant at 0.364 SD, though not statistically significant ($p > 0.10$; β_1 in Column 1). One possible explanation for the difference is that boy-biased families might have invested more in the reading of boys than other families. Second, the intervention had no statistically significant effect on the reading scores of girls in either of the two families (0.069 SD and − 0.031 SD, $p > 0.10$; β_2 in column 1 and 2). However, the effect on boys from boy-biased families was substantially greater than on those from other family types (− 0.230 SD versus − 0.068 SD; $\beta_2 + \beta_3$ in columns 1 and 2), with a seemingly unrelated test yielding a p value of 0.06. Finally, the intervention effect on the gender gap was significantly larger in families with boy-biased preferences compared to other families (− 0.299 SD versus − 0.037 SD; β_3 in columns 1 and 2), with a seemingly unrelated test p value of 0.02.[12] These findings suggest that students from families with biased gender role beliefs are more susceptible to gender stereotypes.

[11] As a robustness check, we further focused on students whose families have identifiable son preferences (boy-biased families versus non-boy-biased families) and report results in the Appendix Table A8–A10. The results are consistent both qualitatively and quantitatively.

[12] As a robustness check, we also employed a triple-difference regression model to formally test the differential impacts. The results, presented in Appendix Table A11, are consistent with those from the seemingly unrelated tests. For instance, δ_7 in Appendix Table A11, which captures the differential impacts on the gender gap, aligns closely with the findings from the seemingly unrelated tests.

Table 5 Heterogeneity analysis at the family level: son preference and family SES

	Son preference			Family education level			Family income level		
	Boy-biased families	Other families	Boy-biased vs. other families: *p* value	Low-education families	High-education families	Low- vs. high-education families: *p* value	Low-income families	High-income families	Low- vs. high-income families: *p* value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male (β_1)	0.364	− 0.083*	0.24	− 0.067	− 0.031	0.63	− 0.064	− 0.050	0.84
	(0.403)	(0.047)		(0.058)	(0.054)		(0.046)	(0.066)	
Treatment (β_2)	0.069	− 0.031	0.18	− 0.053	0.023	0.33	− 0.022	0.000	0.77
	(0.077)	(0.043)		(0.059)	(0.054)		(0.047)	(0.070)	
Male × Treatment (β_3)	− 0.299***	− 0.037	0.02	− 0.019	− 0.134*	0.29	− 0.062	− 0.073	0.91
	(0.108)	(0.063)		(0.084)	(0.074)		(0.070)	(0.089)	
Constant (α)	− 0.674	− 0.417***		− 0.636***	− 0.131**		− 0.450***	− 0.359***	
	(0.420)	(0.048)		(0.052)	(0.061)		(0.043)	(0.095)	
Observations	1,179	3,543		2,490	2,232		3,352	1,370	
R-squared	0.400	0.313		0.337	0.285		0.328	0.305	
Class fixed effects	Yes	Yes		Yes	Yes		Yes	Yes	
Controls for family characteristics	Yes	Yes		Yes	Yes		Yes	Yes	
Controls for student characteristics	Yes	Yes		Yes	Yes		Yes	Yes	
F-test of coefficients of interest									
$\beta_2 + \beta_3$	− 0.230***	− 0.068	0.06	− 0.072	− 0.110**	0.60	− 0.083*	− 0.073	0.89
	(0.079)	(0.043)		(0.054)	(0.054)		(0.045)	(0.063)	

Notes: Dependent variables are normalized total scores (SD). To formally test the differential impacts, we perform the seemingly unrelated test and report the *p* values for regression coefficients in columns 3, 6, and 9. Controls for family characteristics include the number of siblings, birth order, high-education families (highest education level of family members is above high school) (yes = 1), high-income families (annual income per capita > 100,000 yuan) (yes = 1), rural hukou (yes = 1), and student received preschool education (yes = 1). Controls for student characteristics include Chinese language test performance last semester (A = 1) and math test performance last semester (A = 1). Standard errors are clustered at the class level and reported in parentheses

Inference: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Besides, although we did not find significant heterogeneous effects when examining family SES alone (columns 4 to 9 in Table 5), a deeper analysis reveals the interaction between son preference and family SES (Tables 6 and 7). First, in the absence of intervention, boys in low-education and boy-biased families performed significantly better in reading than girls (0.831 SD; β_1 in column 1 of Table 6). Similarly, boys in low-income and boy-biased families outperformed girls (0.290 SD; β_1 in column 1 of Table 7), suggesting that these families may have invested more in boys' reading abilities. Second, although the intervention had negligible effects on girls across all family types (β_2 in Tables 6 and 7), the negative impact on boys occurred mainly among boys from boy-biased and low-SES families. Specifically, the intervention effects were −0.198 SD ($\beta_2 + \beta_3$ in column 1 of Table 6) for boys from low-education, boy-biased families and −0.198 SD ($\beta_2 + \beta_3$ in column 1 of Table 7) for those from low-income, boy-biased families. Finally, in a similar pattern, the intervention's impact on the gender gap was most pronounced in boy-biased and low-SES families (β_3 in column 1 of Tables 6 and 7). These findings suggest that gender stereotyping may be most deeply rooted in boy-biased and low-SES families.

Overall, the results suggest a statistically significant heterogeneous effect between boys from boy-biased families and boys from non-boy-biased families, especially among low-SES families. On the one hand, our results provide empirical evidence for the hypothesis that variation in educational gender gaps between families with different SES could be due to variations in gender stereotypes in those families. This hypothesis was proposed to help explain the dynamics between gender differences in education and family SES; however, to our knowledge, it has not been empirically tested before our study (Penner and Paret 2008; Cobb-Clark and Moschion 2017). On the other hand, our findings also imply that low levels of education and income play important roles in the formation and intergenerational transmission of gender stereotypes within the family.

5.2 Heterogeneous effects of the intervention on students with male and senior Chinese language teachers

Next, we examined the heterogeneous effect of the intervention according to the gender and seniority of the Chinese language teachers (Table 8). While previous research suggests that a gender match between teachers and students can influence educational outcomes by providing gender role models, our findings did not reveal significant differential effects of the intervention based on teacher gender (columns 1 to 3 in Table 8). However, we observed a statistically significant heterogeneous effect that is consistent with our hypothesis about teachers' experiences. Specifically, the results show that in classrooms with a senior teacher, the intervention has a negative and statistically significant effect on the reading performance of boys (− 0.168 SD, $p < 0.01$; $\beta_2 + \beta_3$ in column 4 of Table 8). In contrast, the effect was not statistically significant in classrooms with a junior teacher (− 0.044 SD, $p > 0.10$; $\beta_2 + \beta_3$ in column 5 of Table 8), with a seemingly unrelated test p value of 0.08. The effects of the intervention on the reading performance of girls were statistically insignificant in the two regressions (β_2

Table 6 Heterogeneity analysis at the family level: interaction between son preference and family education level

	Son preference and family education level				*p* values		
	Low-education and boy-biased families	Low-education and other families	High-education and boy-biased families	High-education and other families	(2) vs. (1): *p* value	(3) vs. (1): *p* value	(4) vs. (1): *p* value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Male (β_1)	0.831** (0.347)	− 0.088 (0.071)	− 0.533 (0.384)	− 0.064 (0.063)	0.00	0.00	0.00
Treatment (β_2)	− 0.028 (0.122)	− 0.072 (0.060)	0.096 (0.118)	0.003 (0.065)	0.70	0.42	0.80
Male × treatment (β_3)	− 0.170 (0.149)	0.045 (0.096)	− 0.307* (0.168)	− 0.108 (0.084)	0.15	0.49	0.70
Constant (α)	− 1.342*** (0.351)	− 0.695*** (0.062)	0.587 (0.381)	− 0.165** (0.067)			
Observations	722	1768	457	1775			
R-squared	0.432	0.352	0.483	0.280			
Class fixed effects	Yes	Yes	Yes	Yes			
Controls for family characteristics	Yes	Yes	Yes	Yes			
Controls for student characteristics	Yes	Yes	Yes	Yes			
F-test of coefficients of interest							
$\beta_2 + \beta_3$	− 0.198*	− 0.027	− 0.211	− 0.105*	0.15	0.93	0.40
	(0.103)	(0.070)	(0.136)	(0.057)			

Notes: Dependent variables are normalized total scores (SD). To formally test the differential impacts, we perform the seemingly unrelated test and report the *p* values for regression coefficients in columns 5, 6, and 7. Controls for family characteristics include the number of siblings, birth order, high-education families (highest education level of family members is above high school) (yes = 1), high-income families (annual income per capita > 100,000 yuan) (yes = 1), rural hukou (yes = 1), and student received preschool education (yes = 1). Controls for student characteristics include Chinese language test performance last semester ($A = 1$) and math test performance last semester ($A = 1$). Standard errors are clustered at the class level and reported in parentheses

Inference: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 7 Heterogeneity analysis at the family level: interaction between son preference and family income level

	Son preference and family income level				*p* values		
	Low-income and boy-biased families	Low-income and other families	High-income and boy-biased families	High-income and other families	(2) vs. (1): *p* value	(3) vs. (1): *p* value	(4) vs. (1): *p* value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Male (β_1)	0.290 (0.439)	− 0.098* (0.053)	− 0.277 (0.330)	− 0.082 (0.077)	0.34	0.24	0.36
Treatment (β_2)	0.075 (0.094)	− 0.062 (0.051)	0.068 (0.166)	0.011 (0.090)	0.14	0.97	0.56
Male × treatment (β_3)	− 0.274** (0.122)	0.013 (0.079)	− 0.286 (0.372)	− 0.077 (0.106)	0.03	0.97	0.16
Constant (α)	− 0.706 (0.456)	− 0.424*** (0.053)	0.364 (0.415)	− 0.482*** (0.110)			
Observations	914	2,438	265	1,105			
R-squared	0.432	0.333	0.549	0.325			
Class fixed effects	Yes	Yes	Yes	Yes			
Controls for family characteristics	Yes	Yes	Yes	Yes			
Controls for student characteristics	Yes	Yes	Yes	Yes			
F-test of coefficients of interest							
$\beta_2 + \beta_3$	− 0.198** (0.092)	− 0.050 (0.056)	− 0.218 (0.293)	− 0.066 (0.067)	0.14	0.94	0.21

Notes: Dependent variables are normalized total scores (SD). To formally test the differential impacts, we perform the seemingly unrelated test and report the *p* values for regression coefficients in columns 5, 6, and 7. Controls for family characteristics include the number of siblings, birth order, high-education families (highest education level of family members is above high school) (yes = 1), high-income families (annual income per capita > 100,000 yuan) (yes = 1), rural hukou (yes = 1), and student received preschool education (yes = 1). Controls for student characteristics include Chinese language test performance last semester ($A = 1$) and math test performance last semester ($A = 1$). Standard errors are clustered at the class level and reported in parentheses

Inference: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

in columns 4 and 5 of Table 8). Consequently, the intervention had a more pronounced impact on the gender gap in classrooms with a senior Chinese language teacher (− 0.230 SD versus 0.016 SD; β_3 in columns 4 and 5 of Table 8), with a seemingly unrelated test p value of 0.02.

5.3 Heterogeneous effects of the intervention on students with rural cultural characteristics

Finally, we explored whether Chinese traditional farming culture moderates the effect of interventions on gender stereotypes (Table 9). Our results suggest that there are no significant heterogeneous effects. The original gender gap (β_1), as well as the intervention's impact on girls (β_2), boys ($\beta_2 + \beta_3$), and the gender gap (β_3), were comparable between students with urban and rural hukou. One explanation for this result may be the frequent migration between rural and urban areas and converging beliefs about gender roles.

6 Conclusions

Using data from a randomized controlled trial of 5224 fifth-grade students in East China, this paper provides a novel test for the hypothesis that evoking a gender stereotype creates gender gaps in education through self-fulfilling prophecies. We focus on student reading performance, in which boys are the stereotyped gender, which has been surprisingly overlooked in previous stereotype threat literature. Students in the treatment group were exposed to the intervention of evoking a gender stereotype, indicating the expected outperformance of girls over boys in reading, while those in the control group were not exposed to any stereotype threat. Relying on a randomized controlled trial to analyze this issue enables us to overcome any reverse causality bias and identify the underlying causal relationships.

Our main findings are threefold. First, in the control group, boys performed worse than girls without any intervention. Second, the randomized controlled trial reveals that the stereotype intervention reduced the reading scores of boys while leaving girls unaffected, thus serving to perpetuate the gender gap in education (at least in the case of reading). We explored potential mechanisms and identified increased anxiety as the primary factor contributing to the stereotype effect. Third, our heterogeneity analysis shows that students from environments with biased gender role beliefs—such as low-SES and boy-biased families, or classes with senior teachers—were more susceptible to the impact of gender stereotypes.

Although we do not find a statistically significant effect of the intervention on the overall gender gap, the economic significance of our findings remains noteworthy. One possibility is that our stereotype nudging approach was too simple to produce a more pronounced (statistically significant) effect. However, the fact that the intervention significantly lowers the reading performance of boys while not affecting girls helps to relieve this concern. More importantly, unlike our one-time intervention, real-life stereotype nudging often occurs implicitly and repeatedly throughout the development of

Table 8 Heterogeneity analysis at the class level: gender and seniority of the Chinese language teacher

	Teacher gender			Teacher seniority		
	Female Chinese language teachers	Male Chinese language teachers	Female vs. male Chinese language teachers: *p* value	Chinese language teachers aged 40 and above	Chinese language teachers aged below 40	Chinese language teachers aged 40 and above vs. those aged below 40: *p* value
	(1)	(2)	(3)	(4)	(5)	(6)
Male (β_1)	− 0.024 (0.044)	− 0.202* (0.104)	0.10	0.004 (0.055)	− 0.086 (0.060)	0.26
Treatment (β_2)	− 0.002 (0.044)	− 0.063 (0.079)	0.48	0.062 (0.054)	− 0.060 (0.056)	0.11
Male × treatment (β_3)	− 0.090 (0.064)	− 0.087 (0.091)	0.97	− 0.230*** (0.071)	0.016 (0.083)	0.02
Constant (α)	− 0.426*** (0.042)	− 0.046 (0.143)		− 0.452*** (0.062)	0.318*** (0.058)	
Observations	4,150	572		1,947	2,775	
R-squared	0.278	0.490		0.350	0.283	
Class fixed effects	Yes	Yes		Yes	Yes	
Controls for family characteristics	Yes	Yes		Yes	Yes	
Controls for student characteristics	Yes	Yes		Yes	Yes	
F-test of coefficients of interest						
$\beta_2 + \beta_3$	− 0.092** (0.041)	− 0.150* (0.071)	0.46	− 0.168*** (0.050)	− 0.044 (0.054)	0.08

Notes: Dependent variables are normalized total scores (SD). To formally test the differential impacts, we perform the seemingly unrelated test and report the *p* values for regression coefficients in columns 3 and 6. Controls for family characteristics include the number of siblings, birth order, high-education families (highest education level of family members is above high school) (yes = 1), high-income families (annual income per capita > 100,000 yuan) (yes = 1), rural hukou (yes = 1), and student received preschool education (yes = 1). Controls for student characteristics include Chinese language test performance last semester ($A = 1$) and math test performance last semester ($A = 1$). Standard errors are clustered at the class level and reported in parentheses

Inference: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 9 Heterogeneity analysis at the cultural level: rural hukou

	Hukou type		
	Urban hukou	Rural hukou	Urban versus rural hukou: *p* value
	(1)	(2)	(3)
Male (β_1)	− 0.069 (0.052)	− 0.047 (0.067)	0.78
Treatment (β_2)	− 0.014 (0.049)	− 0.017 (0.057)	0.96
Male × treatment (β_3)	− 0.070 (0.075)	− 0.071 (0.087)	0.99
Constant (α)	− 0.321*** (0.061)	− 0.829*** (0.070)	
Observations	2790	1932	
R-squared	0.297	0.335	
Class fixed effects	Yes	Yes	
Controls for family characteristics	Yes	Yes	
Controls for student characteristics	Yes	Yes	
F-test of coefficients of interest			
$\beta_2 + \beta_3$	− 0.083 (0.051)	− 0.089 (0.061)	0.95

Notes: Dependent variables are normalized total scores (SD). To formally test the differential impacts, we perform the seemingly unrelated test and report the p-values for regression coefficients in columns 3. Controls for family characteristics include the number of siblings, birth order, high-education families (highest education level of family members is above high school) (yes = 1), high-income families (annual income per capita > 100,000 yuan) (yes = 1), rural hukou (yes = 1), and student received preschool education (yes = 1). Controls for student characteristics include Chinese language test performance last semester ($A = 1$) and math test performance last semester ($A = 1$). Standard errors are clustered at the class level and reported in parentheses

Inference: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

children, driven by families, teachers, and societal norms—even without full awareness. Such ongoing reinforcement could have a deeply concerning and long-lasting effect.

Despite these findings, our study has two main limitations. First, the nudge we employed was based on a straightforward mention of gender differences in test scores. While this method mirrors real-world experiences where gender stereotypes are presented as general information, more personalized nudges tailored to individual students may yield stronger outcomes. Future research, particularly those focused on directly mitigating stereotype threats, could explore whether personalized interventions, adjusted to the specific characteristics and experiences of students, might result in more substantial and lasting effects.

Second, our study lacks follow-up data to assess whether the effects of the stereotype intervention persist beyond the initial assessment. While we identified immediate impacts, we were unable to evaluate their long-term influence on student reading performance. Additionally, the potential for stereotype interventions to affect

subsequent human capital investments, such as the increased focus on math by boys after exposure to reading-related stereotypes, remains unexplored. Although prior research suggests that teacher-held gender stereotypes can have lasting effects on student outcomes (Alan et al. 2018; Lavy and Sand 2018), there is limited causal evidence on the long-term impact of a brief, one-time intervention. Future research should explore the persistence of stereotype interventions over time and examine their broader implications for the long-term human capital development of students.

In conclusion, our study indicates that gender stereotypes can be easily created but produce a powerful effect on students. Any policies aiming to erase gender stereotypes held by families, teachers, and all other people surrounding children could bring substantial value to reducing gender gaps in early educational achievement. However, to our knowledge, few studies have been conducted to determine how to reduce gender stereotypes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00148-025-01102-6.

Acknowledgements We would like to extend our sincere thanks to editor Alfonso Flores-Lagunes and three anonymous referees for their valuable comments and suggestions, which have significantly contributed to improving the quality of our research. We acknowledge the support of local education departments and sample schools in the coordination of the field survey and the teachers and graduate students from Anhui Agricultural University in the data collection. We also want to thank the teachers, students, and their families from the sample classes for their patience in completing the survey.

Funding This work was supported by the Chen Yet-Sen Family Foundation (via the Doers Cultural & Educational Foundation), the China Merchants Foundation, and the China Population Welfare Foundation.

Data Availability Data used for analysis in this paper is available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

References

Alan S, Ertac S, Mumcu I (2018) Gender stereotypes in the classroom and effects on achievement. Rev Econ Stat 100(5):876–890

Alesina A, Giuliano P, Nunn N (2013) On the origins of gender roles: women and the plough. Quart J Econ 128(2):469–530

Antecol H, Eren O, Ozbeklik S (2015) The effect of teacher gender on student achievement in primary school. J Law Econ 33(1):63–89

Aronson J, Lustina MJ, Good C, Keough K, Steele CM, Brown J (1999) When white men can't do math: necessary and sufficient factors in stereotype threat. J Exp Soc Psychol 35(1):29–46

Arrow K (1973) The theory of discrimination, O. Ashenfelter and A. Rees, Discrimination in Labor Markets. Princeton, NJ: Princeton University Press, 3–33.

Beilock SL, Gunderson EA, Ramirez G, Levine SC (2010) Female teachers' math anxiety affects girls' math achievement. Proceed Nat Acad Sci 107(5):1860–63

Ben-Zeev T, Fein S, Inzlicht M (2005) Arousal and stereotype threat. J Exp Soc Psychol 41(2):174–181

Bertocchi G, Bozzano M (2016) Women, medieval commerce, and the education gender gap. J Comp Econ 44(3):496–521
Bettinger EP, Long BT (2005) Do faculty serve as role models? the impact of instructor gender on female students. Am Econ Rev 95(2):152–157
Bordalo P, Coffman K, Gennaioli N, Shleifer A (2016) Stereotypes. Quart J Econ 131(4):1753–94
Brenøe AA (2021) Brothers increase women's gender conformity. J Population Econ 1–38.
Carlana M (2019) Implicit stereotypes: evidence from teachers' gender bias. Quart J Econ 134(3):1163–1224
Carrell SE, Page ME, West JE (2010) Sex and science: how professor gender perpetuates the gender gap. Quart J Econ 125(3):1101–44
Chaffee KE, Plante I, Good C, Aronson JM, Simon-Benoît K, Isabelle G (2024) When stereotypes disadvantage boys: strength of stereotypes in mathematics and language arts and their relations with grades. J Appl Soc Psych 54(2):71–82
CNBS. 2021. *China Statistical Yearbook*. China Statistics Press.
Cobb-Clark DA, Moschion J (2017) Gender gaps in early educational achievement. J Popul Econ 30(4):1093–1134
Dahl GB, Moretti E (2008) The demand for sons. Rev Econ Stud 75(4):1085–1120
Danaher K, Crandall CS (2008) Stereotype threat in applied settings re-examined. J Appl Soc Psychol 38(6):1639–1655
Das Gupta M, Zhenghua J, Bohua L, Zhenming X, Chung W, Hwa-Ok B (2003) Why is son preference so persistent in East and South Asia? A cross-country study of China, India and the Republic of Korea. J Dev Stud 40(2):153–87
Dee TS (2007) Teachers and the gender gaps in student achievement. J Human Res 42(3):528–554
Del Carpio L, Guadalupe M (2022) More women in tech? Evidence from a field experiment addressing social identity. Manage Sci 68(5):3196–218
Dossi G, Figlio D, Giuliano P, Sapienza P (2021) Born in the family: preferences for boys and the gender gap in math. J Econ Behav Organization 183:175–88
Evans DK, Yuan F (2022) How big are effect sizes in international education studies? Educ Eval Policy Anal 44(3):532–540
Fitzgerald J, Shanahan T (2000) Reading and writing relations and their development. Educational Psychologist 35(1):39–50
Flore PC, Wicherts JM (2015) Does stereotype threat influence performance of girls in stereotyped domains? A Meta-Analysis J School Psychol 53(1):25–44
Gao Q, Wang H, Mo D, Shi Y, Kenny K, Rozelle S (2018) Can reading programs improve reading skills and academic performance in Rural China? China Econ Rev 52:111–25
Gao Q, Wang H, Chang F, An Q, Yi H, Kenny K, Shi Y (2022) Feeling bad and doing bad: student confidence in reading in Rural China. Compare J Comparative Int Ed 52(2):269–88
Guiso L, Sapienza P, Zingales L (2006) Does culture affect economic outcomes? J Econ Perspectives 20(2):23–48
Huang W, Wang Y, Wu H, Zhou Y (2025) The motherhood penalty and low fertility in China: a pseudo-event study. J Population Econ 38(1):1–29
Istenič MČ (2007) Attitudes towards gender roles and gender role behaviour among urban, rural, and farm populations in Slovenia. J Comp Fam Stud 38(3):477–496
Kraft MA (2023) The effect-size benchmark that matters most: education interventions often fail. Educ Res 52(3):183–187
Lavy V (2008) Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. J Public Econ 92(10–11):2083–2105
Lavy V, Sand E (2018) On the origins of gender gaps in human capital: short-and long-term consequences of teachers' biases. J Public Econ 167:263–279
Lei X, Shen Y, Smith JP, Zhou G (2017) Sibling gender composition's effect on education: evidence from China. J Popul Econ 30(2):569–90
Li J, McLellan R (2021) Is language learning a feminine domain? Examining the content and stereotype threat effect of female-language stereotypes among Efl learners in China. Contemp Educ Psychol 66:101991
Li H, Shi X (2025) The effect of the one-child policy on fertility in China: identification based on difference-in-differences. J Popul Econ 38(1):2
Lily S (1994) Socioeconomic status, parents' sex-role stereotypes, and the gender gap in computing. J Res Comput Educ 26(4):433–451

Lippmann Q, Senik C (2018) Math, girls and socialism. J Comp Econ 46(3):874–888
Liu Z (2005) Institution and inequality: the Hukou system in China. J Comp Econ 33(1):133–157
Machin S, Pekkarinen T (2008) Global sex differences in test score variability. Science 322(5906):1331–1332
Martin MO, Mullis IVS, Hooper M (2017) Methods and procedures in PIRLS 2016. International Association for the Evaluation of Educational Achievement.
McLaughlin MJ, Speirs KE, Shenassa ED (2014) Reading disability and adult attained education and income: evidence from a 30-year longitudinal study of a population-based sample. J Learn Disabil 47(4):374–86
Mullis IVS, Martin MO, Foy P, Drucker KT (2012). PIRLS 2011 International Results in Reading. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Mullis IVS, Martin MO, Foy P, Hooper M (2017) PIRLS 2016 International Results in Reading. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Muntoni F, Retelsdorf J (2018) Gender-specific teacher expectations in reading—the role of teachers' gender stereotypes. Contemp Educ Psychol 54:212–220
Muralidharan K, Sheth K (2016) Bridging education gender gaps in developing countries: the role of female teachers. J Human Res 51(2):269–297
Nguyen H-H, Ryan AM (2008) Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. J Appl Psychol 93(6):1314
Nosek BA, Smyth FL, Sriram N, Lindner NM, Devos T, Ayala A, Bar-Anan Y, Bergh R, Cai H, Gonsalkorale K (2009) National differences in gender–science stereotypes predict national sex differences in science and math achievement. Proceed Nat Acad Sci 106(26):10593–97
OECD (2019) PISA 2018 Results (Volume II): where all students can succeed. OECD Publishing, Paris, PISA
Pansu P, Régner I, Max S, Colé P, Nezlek JB, Huguet P (2016) A burden for the boys: evidence of stereotype threat in boys' reading performance. J Exp Soc Psychol 65:26–30
Penner AM, Paret M (2008) Gender differences in mathematics achievement: exploring the early grades and the extremes. Soc Sci Res 37(1):239–253
Phelps ES (1972) The statistical theory of racism and sexism. Am Econ Rev 62(4):659–661
Rose E (2000) Gender bias, credit constraints and time allocation in Rural India. Econ J 110(465):738–758
Rosenblum D (2013) The effect of fertility decisions on excess female mortality in India. J Popul Econ 26(1):147–180
Savolainen H, Ahonen T, Aro M, Tolvanen A, Holopainen L (2008) Reading comprehension, word reading and spelling as predictors of school achievement and choice of secondary education. Learn Instruct 18(2):201–10
Spencer SJ, Steele CM, Quinn DM (1999) Stereotype threat and women's math performance. J Exp Soc Psychol 35(1):4–28
Spencer SJ, Logel C, Davies PG (2016) Stereotype threat. Annu Rev Psychol 67(1):415–37
Steele CM (1997) A threat in the air: how stereotypes shape intellectual identity and performance. Am Psychol 52(6):613
Steele CM, Aronson J (1995) Stereotype threat and the intellectual test performance of African Americans. J Pers Soc Psychol 69(5):797
Yi H, Mo D, Wang H, Gao Q, Shi Y, Wu P, Abbey C, Rozelle S (2019) Do resources matter? Effects of an in-class library project on student independent reading habits in primary schools in Rural China. Read Res Quart 54(3):383–411
Zhang W, Zou X, Luo C, Yuan L (2024) Hukou reform and labor market outcomes of urban natives in China. J Popul Econ 37(2):50