Inspirational or Demotivational: Experimental Evidence on the Impacts of Role Models* †

Aurelia (Aochun) Di‡

October 2024

Abstract

High-achieving role models can inspire people and raise aspirations, but may negatively affect some individuals who fail to meet their goals. I study what role model characteristics influence their effectiveness in improving academic performance. Using a randomized controlled trial with students across five middle schools in China, I compare the impacts of role models with different success levels. Two months later, students exposed to higher-achieving role models improved test scores by 0.07-0.18 standard deviations on average, whereas those exposed to moderately achieving role models experienced a 28.8% and 26.6% reduction in the likelihood of depression and stress, respectively. Higher-achieving role models improve low-performing girls' academic outcomes but negatively affect their mental health, as these girls invested more effort but still found their improved performance falling short of their elevated aspirations. This paper highlights the negative impacts of role models on mental health as a trade-off for enhancing performance in underperforming subgroups, emphasizing the need to consider mental health when implementing role model interventions.

Keywords: Role models; Mental health; Secondary education; Gender; China.

JEL codes: J16, J24, J70, I25, I21, O12.

*I thank my advisors, Rachel Heath, Emma Riley, and Xu Tan, for their comments, advice, and unreserved support. I thank participants at the University of Washington Labor-Development Brownbag, the University of Washington Political Economy Forum, the 2024 Northwest Development Workshop, and the WEAI 99th Annual Conference for their suggestions. I thank two anonymous role models for their generous sharing. The study has been approved by the University of Washington (STUDY00017191). The experimental design was registered at the AEA Registry (RCT ID: AEARCTR-0010826), with a pre-analysis plan available. Deviations from the pre-analysis plan and the reasons for them are described in Appendix Table D18. All errors are mine.

†This paper will be presented in the 2025 ASSA Annual Meeting.

‡Department of Economics, University of Washington. Email: adi317@uw.edu

1 Introduction

Role models positively influence adolescents by enhancing academic performance and shaping decision-making (Serra, 2022). They model positive behaviors, challenge inaccurate views, and raise aspirations (Beaman et al., 2012; Bernard et al., 2014; Dasgupta and Asgari, 2004; Jensen, 2012; Jensen and Oster, 2009; Lybbert and Wydick, 2018). However, there are two potential concerns about the effectiveness of role models. On the one hand, role models should be successful, as those without notable achievements are unlikely to inspire their role aspirants (Morgenroth et al., 2015). On the other hand, while higher-achieving role models are more inspirational and make goals more desirable, they can negatively impact a subgroup of their role aspirants who may feel frustrated if they fail to reach their elevated goals (Genicot and Ray, 2020).

In this paper, I study whether different role model characteristics enhance or diminish their effectiveness, particularly focusing on the success levels of role models relative to the abilities of their role aspirants. I examine the effects of role models with varying success levels on students' academic performance and mental health. Further, I investigate role aspirants' beliefs, efforts, and aspirations as channels by which role models of different success levels influence these outcomes.

To investigate these questions, I conduct a randomized controlled trial with 1,920 students from 49 classes across 5 middle schools in China. In this region, middle school enrollment and class assignments are independent of students' academic performance. Students were randomized at the class level into three treatment groups or one control group. The treatment groups feature *Very Successful & Similar-Background (Very Successful)* role models, *Very Successful & General-Background (VS-General)* role models, and *Moderately Successful & Similar-Background (Moderately Successful)* role models, respectively. In the same week as the treatments, students in the control group participated in a non-academic class meeting.

To explore what role model characteristics determine their effectiveness, I ensure that all the other aspects of role models remain constant. I use the same interviews with the same

role models across all the treatments, while selectively and strategically revealing their characteristics in each treatment. I invite one female and one male role model to minimize the gender-related effects of role models. Both role models share backgrounds similar to those of students and have accomplished exceptional academic achievements[1]. To avoid direct interaction, the role models were interviewed separately using online meeting software, and their interviews were directly recorded by the software. Each treatment video is approximately 29 minutes long, starting with a 4-minute introduction of the two role models, which differs across the three treatment groups. The remaining 25 minutes are identical, where the role models share their effective learning strategies and discuss gender.

In the Very Successful role model treatment, the video fully reveals the achievements of these two role models, especially those during their secondary education, and highlights their background similarities shared with the students. In the VS-General role model treatment, the role models' achievements are fully revealed, yet background similarities are not explicitly mentioned. In the Moderately Successful role model treatment, background similarities are disclosed as in the first treatment, but role models' achievements are described less precisely to create the perception of "*moderate success*", *e.g.*, stating the role models were among the top 50 rather than the top 2 students.

My study finds a trade-off due to exposure to role models of different success levels. The Very Successful role models enhance academic performance, increasing the midterm total test score by 0.073 standard deviations, measured within one month after the interventions. This effect becomes stronger in the final exam, conducted around 2 months later, resulting in an increase of 0.085 standard deviations in the final total test score. While their impact on the midterm math test is insignificant, the Very Successful role models lead to an improvement of 0.184 standard deviations in the final math exam, suggesting that certain skills take time

[1]Both role models attended the top senior high school in the city via taking the Senior High School Entrance Examination (admission rate in 2023: 0.97%). Later, they took the National College Entrance Examination and pursued their undergraduate education at Peking or Tsinghua University (admission rate of the province in 2023: 0.128%). 97% of the students in my study aspired to take these exams after completing middle school.

to develop.[2] In contrast, the Moderately Successful role models do not affect test scores, and the differences in the effects of the Moderately Successful and the Very Successful role models are large and statistically significant.

My sample has a high rate of poor mental health, with 41.6% of students reporting depression in the past two weeks and 34.9% experiencing heavy stress in the control group. I find that the Moderately Successful role models effectively improve mental health, leading to a 28.8% decrease in the probability of depression and a 26.6% reduction in the probability of stress. The Very Successful role models do not affect mental health on average, and the gap between the impacts of the Very Successful role models and the Moderately Successful role models is found to be statistically significant and large.

Both girls and boys experienced this trade-off, as the Very Successful role models enhance their academic performance, whereas the Moderately Successful role models improve their mental health. Girls were more likely to face mental health issues than boys, with around 51.1% of girls reporting depression in the past two weeks compared to only 32.9% of boys. The impacts of the Very Successful role models on girls' mental health are significantly different from the mental health relief provided by the Moderately Successful role models.

I analyze the treatment effects across four quartiles of students' baseline ability. I find that bottom-performing (the lowest quartile) girls, after exposure to the Very Successful role models, significantly improved their midterm total test score by 0.200 standard deviations but also experienced a significant increase in their poor mental health index by 0.385 standard deviations. The positive impact on test scores declined in the final exam, with an improvement of 0.149 standard deviations among the bottom-performing girls. These findings indicate that these bottom-performing girls benefit from the higher-achieving role models in improving their academic performance, but it comes with a cost to their mental health. To examine the multiple baseline characteristics of those who experienced the greatest treatment effects, I use a machine learning approach – *casual forest* – developed by

[2]The positive effects of the Very Successful role models on academic outcomes are persistent and observed in the long-term Senior High School Entrance Examination.

Wager and Athey (2018). Results confirm that the Very Successful role models have trade-off effects on bottom-performing girls.

To understand why a subgroup of role aspirants have improved their academic outcomes but still experienced poor mental health, I investigate the potential mechanisms by which the Very Successful role models affect the girls. I find that the Very Successful role models encourage bottom-performing girls to adopt higher aspirations for their Senior High School Entrance Examination and increase their learning hours. Although the Very Successful role models improve the first post-intervention exam of bottom-performing girls, they do not affect their exam rankings nor help these girls to move out of the bottom quartile. Thus, the mental health decline in this subgroup of role aspirants is attributed to their improved academic performance falling short of their elevated aspirations.

This paper contributes to the literature by incorporating mental health in evaluating role model interventions. Mental health relief is crucial for the development of students in the educational context, as mental health issues can raise many concerns, including higher likelihood of school dropout (Shi et al., 2015), poor academic performance (Kötter et al., 2017; Wang et al., 2015), poor sleep (Bernert et al., 2007), and poor physical health (Stults-Kolehmainen and Sinha, 2014). This paper addresses students' mental health problems through moderately achieving role models.

This paper also contributes to the role model literature by studying role model characteristics, particularly focusing on their success levels. Theoretical studies (Morgenroth et al., 2015)[3] conclude that role models without significant achievements fail to influence their role aspirants effectively, which is unexplored by the empirical literature. This paper addresses this gap, finding that while moderately achieving role models do not significantly boost academic performance, they significantly reduce the likelihood of experiencing mental health issues. In addition, to ensure all the other factors of role models are constant, this paper

[3]Morgenroth et al. (2015) concludes that role model impacts follow an inverted U-shape curve as the prominence of role models increases – too successful role models may fail to persuade the role aspirants that the goals are attainable to them, whereas role models lacking notable achievements would not inspire at all.

also introduces a new intervention design. Each treatment uses the same interviews with the same role models, with the treatment video edited to selectively and strategically reveal specific information without altering the rest.

Despite the positive impacts of role models[4], existing literature has overlooked the potential downsides of role modeling. This paper fills this gap by showing that higher-achieving role models negatively affect mental health of underperforming girls as a trade-off for enhancing their academic outcomes. The mental health issues arise because these girls' improved educational performance still falls short of their elevated aspirations. These findings connect to another body of research, which demonstrates that failing to achieve goals leads to negative mental health outcomes (Genicot and Ray, 2017, 2020).

Leveraging cost-effective treatments, this paper emphasizes the importance of considering and supporting mental health in underperforming subgroups when implementing role model interventions. It also provides policy implications for improving educational attainment in China. To address the low educational attainment rate in China (Khor et al., 2016), my paper suggests that schools could invite high-achieving role models to boost students' educational outcomes. 74% of rural students in China are at risk for mental health issues, which are correlated with high dropout rates (Wang et al., 2015). My paper tackles this issue by suggesting schools invite moderately achieving role models to improve students' mental health. Scaling up these role model interventions by encouraging schools to invite their successful alumni can further enhance the cost-effectiveness of this approach.

The rest of the paper is assigned as follows. Section 2 provides the theoretical framework. Section 3 discusses the background and experiment design. Section 4 describes the data and empirical analysis strategies. Section 5 provides the main findings. Section 6 discusses the policy implications. Section 7 concludes. Appendices follow the reference list.

[4]Numerous studies show that role models positively influence decision-making and performance (see Serra, 2022 for a review). Beyond the educational context (Breda et al., 2023; Di, 2024; Golan and You, 2021; Kipchumba et al., 2021; Riley, 2019), role models also positively a wide range of economic outcomes, such as motivating inexperienced entrepreneurs to develop their businesses (Lafortune et al., 2018) and empowering HIV-positive women to invest in income- and welfare-generating activities (Lubega et al., 2021).

2 Theoretical framework

The baseline survey suggests that approximately 97% of the students in my sample plan to participate in the Senior High School Entrance Examination and apply to upper secondary schools based on their exam results and school preference.

To make predictions about the impacts of the interventions, I construct a model based on Genicot and Ray (2017, 2020). The model is premised on the idea that realizing aspirations in the next period increases the current utility, yet comes at the cost of more learning effort spent in the current period and a potential cost of not reaching the goals in the next period.

The model has two periods. In the Period 1, a student achieves a total test score of y_1 in her most recent exam. I assume that the student performed to her usual standards during the most recent exam, such that her total test score reflects her actual ability. The student aspires to achieve a total test score α in the next exam and invests effort e. If the student does not have an aspiration, then her $\alpha = 0$. Thus, the most recent total test score y_1 also suggests the "distance" between the student's current position and her aspiration α. The larger this distance is, the more effort e the student needs to increase her chance of achieving her aspiration α. The utility function in Period 1 can be written as:

$$U_1(y_1, e) = u_1(y_1) - c(e), \tag{1}$$

where $u_1(\cdot)$ is the utility from current achievement and $c(\cdot)$ is the cost of effort that satisfies $c'(e) > 0$ and $c''(e) < 0$. I assume $u_1(y_1)$ to be a constant $\bar{u}_1$ which is not influenced by e.

In the second period, the student takes an exam and achieves a total test score of y_2. I assume no additional effort in Period 2 for simplicity. Assume effort e in Period 1 will translate into $g(e)$ additional score in Period 2 with an uncertainty term ε. That means $y_2 = y_1 + g(e) + \varepsilon$, where $g(\cdot)$ satisfies $g'(e) > 0$ and $g''(e) < 0$ and $\varepsilon \sim N(0, \sigma^2)$. If y_2 does not meet the aspiration α in Period 1, then the student feels frustrated, which can be captured by a frustration function $\psi(\alpha - y_2)$. I formulate the student's utility function in

Period 2 as follows:

$$E[U_2(y_2)] = E[u_2(y_2)] - E[\psi(\alpha - y_2)], \text{ where } \psi(\alpha - y_2) = \begin{cases} 0 & \text{if } y_2 \geq \alpha \\ \lambda(\alpha - y_2) & \text{if } y_2 < \alpha \end{cases} \tag{2}$$

where $u_2(\cdot)$ is the utility from test score y_2 in Period 2, and $\lambda > 0$ indicating the frustration intensity. For simplicity, I assume that $u_2(y_2) = y_2$.

Assume the student has a discount factor $\beta \in (0, 1]$. Then, the student needs to optimize her utility by solving

$$\max_e E[U] = U_1(y_1, e) + \beta E[U_2(z)] = [\bar{u}_1 - c(e)] + \beta\big\{E[u_2(y_2)] - E[\psi(\alpha - y_2)]\big\},$$

which can be simplified as

$$\max_e \big\{ - c(e) + \beta[y_1 + g(e)] - \beta\lambda \int_{-\infty}^{\alpha} (\alpha - y_2) f(y_2) dy_2 \big\}. \tag{3}$$

Then, the solution to the student's optimization problem is:[5]

$$\beta[1 + \lambda Pr(y_2 < \alpha)]g'(e) = c'(e). \tag{4}$$

In Equation 4, the term $Pr(y_2 < \alpha)$ indicates the probability of the student not achieving her aspiration in Period 2. $\beta\lambda g'(e)$ reflects the marginal disutility from frustration when the aspiration is not met, which increases the marginal benefit of effort. It indicates that the student's marginal cost of effort in Period 1, $c'(e)$, is larger than her marginal benefit of effort in Period 2, $\beta g'(e)$, if she cannot meet her aspiration.

[5]The proof of this equilibrium condition can be found in Appendix Section E.

2.1 How the treatments could affect the students

The role model treatments are integrated into the model by changing aspirations, learning effort, or both, to improve the total test score in the second period. The effectiveness of the treatments may vary depending on the characteristics of role models. Revealing these role model characteristics to students may influence how they interpret the content shared or conveyed by role models, with potentially significant influences on specific subgroups of role aspirants. I broadly classify these mechanisms as aspirations, mental health, stereotype overcoming, and returns to schooling.

Aspirations Role models raise people's aspirations (Beaman et al., 2012; Bernard et al., 2019).The effects of role models on aspirations are observed post-treatment (Bernard et al., 2014) shortly and persist over the long term (Bernard et al., 2023). Theoretically, role models can encourage individuals to *adopt new or higher goals* and *confirm the attainability of existing goals* (Morgenroth et al., 2015).

Higher-achieving role models are expected to be more inspirational to students, encouraging them to adopt higher educational aspiration α. Students might under- or overestimate their probability of achieving their aspirations, *i.e.*, $Pr(y_2 \geq \alpha)$. Moderately successful role models could help these students to confirm whether their existing aspirations are feasible, potentially allowing them to aim higher if they believe higher aspirations are attainable for them too. Empirically, the impacts on aspirations may be more pronounced among underperforming students, because exam-related aspirations are constrained by an upper limit[6] that makes it hard for top performers' aspirations to rise further. Additionally, raised aspirations are expected to accompany an increased effort e to achieve the elevated goals.

Mental health Mental health problems can arise in individuals who fail to realize their aspirations (Genicot and Ray, 2017, 2020). Therefore, although higher-achieving role models are expected to raise aspiration α and increase effort e, they can also negatively affect the

[6]For instance, aspired total test score cannot exceed the maximum possible score; and, aspired exam rankings cannot be higher than ranking number 1.

mental health of students, particularly those who fall short of their aspirations α. On the other hand, if moderately achieving role models can help the students confirm their attainability in realizing their goals, they might improve students' mental health.

Stereotype overcoming Role models challenge stereotypes (Dasgupta and Asgari, 2004; Stout et al., 2011). They improve girls' performance in stereotype-dominating subjects (Di, 2024) and encourage them to pursue STEM majors (Agurto et al., 2021; Breda et al., 2023) through changing their beliefs on those stereotypes. Higher achieving role models are expected to be more effective in counteracting these stereotypes and thus encouraging girls to improve their math performance. I expect stronger impacts among high-performing girls (Agurto et al., 2021), as both they and higher-achieving role models were top performers at school, sharing similarities that enhance the role model impacts (Nguyen, 2008). In contrast, underperforming girls may need additional resources, such as effective learning strategies and time, to challenge these stereotypes and increase self-concept (Di, 2024).

Returns to schooling The achievements of role models can help students estimate the returns to education. This impact is expected to be more prominent to students with a similar socio-economic background to the role models (Nguyen, 2008), as shared backgrounds allows for a more accurate estimate of potential returns. Higher-achieving role models suggest greater returns to schooling, which may increase students' learning effort e (Jensen, 2012) and improve their educational outcomes. Perceptions of higher returns to education can also raise the educational aspiration α, but overly ambitious aspirations may also increase the probability of not achieving aspirations and thus increase the marginal disutility from frustration.

2.2 Model predictions

From the discussion in Section 2.1, the model allows me to make three empirical predictions of the impacts of exposure to role models of varying success levels, which are testable with my experiment. First, in Period 2, both the stereotype-dominating math test score and

the total test score will be higher after exposure to higher-achieving role models. Second, mental health will be improved after exposure to moderately achieving role models. Third, higher-achieving role models will raise mental health issues for students who have invested more effort e in Period 1, yet still find that their improved academic outcomes y_2 are far from reaching their elevated aspiration α.

3 Background and Experiment Design

3.1 Middle School Education in China

The middle school education in China takes three years to complete, spanning from Grade 7 to Grade 9. Middle school students study subjects including Chinese literature, Math, English, Physics, Chemistry, Morality and law, History, Geography, Physical education, Music, Art, and Technology. With few exceptions, middle school students learn Physics in Grades 8-9 and learn Chemistry in Grade 9. The curriculum is standardized across the country, ensuring a consistent level of education for all students.

Upon finishing Grade 9, students can choose whether or not to take the Senior High School Entrance Examination, also known as the "*Zhong-Kao*". This highly competitive exam assesses students' knowledge and skills in various subjects[7] and is critical for students who are interested in upper secondary education. Their results in *Zhong-Kao* are the primary factor in deciding what upper secondary schools students can attend.[8] High scores can secure admission to prestigious senior high schools, which are often seen as a pathway to top universities.

[7]The subjects tested in the Senior High School Entrance Examination are different across cities. In the city where this experiment was conducted, they test Chinese language, Math, English, Physics, Chemistry, Morality and law, History, Geography, and Physical education and health.

[8]One computerized system records all information to match students with upper secondary schools. Information includes students' Entrance Examination scores, their school preferences, and the capacities of those schools.

3.2 The Five Middle Schools

This experiment works with five middle schools, which are located in one low-income[9] district within a low-income province[10]. These 5 middle schools are close to each other geographically. The linear distance is 4.7km on average between each pair of middle schools. The driving distance between any two of the schools is 6.6 km on average, which is equivalent to a 10.9-minute drive[11].

In this location, both middle school enrollment and class assignment follow a randomized process and thus are irrelevant to students' academic performance. Students are enrolled in middle schools based on the address listed on their national ID. If a student's address is qualified for multiple middle schools, then the student will be randomly assigned to one of the eligible schools. Once enrolled, middle school students are randomly assigned to classes[12]. These middle schools do not provide any advanced classes and follow the national standardized curriculum. These schools provide Physics courses for Grades 8-9 students and provide Chemistry courses for Grade 9 students.

3.3 Sample

The sample of this study consists of Grades 7 and 8 students from these 5 middle schools. The baseline student survey shows that 96.68% of surveyed students aspired to pursue upper secondary education in a local public school. This suggests that most students in my experiment planned to participate in the Senior High School Entrance Examination (*Zhong-Kao*) after completing middle school, and they aimed for high scores to gain admission to desirable upper secondary schools. As discussed later, the two role models also followed this path and

[9]This district contributed only 4.4% of the city's GDP from January 2022 to September 2022 [Data source: Open data of the city's official website]

[10]This province has annual GDP falling below the 17% percentile of the province-level GDP in mainland China [Data source: National Bureau of Statistics, GDP by province, years 2019-2022].

[11]These estimates were obtained using Baidu Map, based on real-time data from departures between 5:27-5:55 AM on a Sunday when traffic was light.

[12]In principle, enrolled students will take all the courses and do all the activities within the same class. They will not transfer to another class or a different school until graduation.

achieved great success in their Senior High School Entrance Examination.

A considerable proportion of middle school students face mental health issues in China.[13] My sample suggests that students in my experiment may be at higher risk of experiencing mental health difficulties, with 41.6% of the students in the control group reporting depression and 34.9% feeling stressed in the past two weeks. These students were enrolled in middle schools in 2020 or 2021 during the COVID-19 pandemic, and consequently, they rarely attended school in person. In 2022, when the pandemic was temporarily under control, the middle schools were instructed to resume in-person classes. However, the situation quickly worsened again, leading to local pandemic outbreaks that forced students back into remote learning. This cycle of returning to school and then shifting back to remote learning occurred twice for Grade 7 students and three times for Grade 8 students before the local government announced that all students would continue learning remotely until the start of the Spring 2023 semester.

3.4 The Role Models

To minimize any effects related to the gender of role models, I invited one female and one male role model to participate in this study voluntarily.

The role models share similar backgrounds with the middle school students in my study. They were born and raised in the same city as the students and attended a different[14] middle school located near[15] the five schools recruited for this study. During the interview, the role models described a typical school day from their middle school years, including the landmarks and the street with food and entertainment options near the schools.

Both role models completed their middle school education around twelve years ago, mak-

[13] According to 2013-2014 China Education Panel Survey (CEPS), a nationally representative survey of Chinese middle school students, 31% of the surveyed students reported experiencing depression, 32% reported sadness, and 43% reported feeling unhappy in the past 7 days.

[14] Thus, it is unlikely that the class homeroom teachers in this study have known the role models in person.

[15] The linear distance between the role models' middle school and any of the five middle schools in my experiment is 3.5 km on average, with an estimated driving time of 9.4 minutes on average. The driving time was obtained using Baidu Map, based on real-time data from departures between 5:27-5:55 AM on a Sunday when traffic was light.

ing their experiences relevant and helpful to the current students. They took the Senior High School Entrance Examination to attend a senior high school in the city and later the National College Entrance Examination (*Gao-Kao*) to enroll in a university in mainland China. Most students in my study aspired to take these exams after completing middle school. Thus, this academic path also reflects the aspirations of most students in my study.

Beyond background similarities, the role models have achieved outstanding success among those who have grown up and studied in the same location, as recognized by the students in the study and supported by school administrative statistics. The role models went to the best senior high school[16] in the city. After completing senior high school, the role models maintained exceptional academic performance and pursued their undergraduate education at Peking University or Tsinghua University – the best universities in mainland China[17] – through their extraordinary performance in the National College Entrance Examination.

3.5 Role Model Interviews

To avoid effects from direct interactions among the role models or between the role models and students, each role model was interviewed separately using Zoom or Tencent Meeting[18]. These interviews were recorded directly by the software. Both role models were given the same list of questions before their interviews to ensure consistency. The interviewer did not guide or suggest any answers, ensuring that the responses from the role models were entirely their own. I used Adobe Premiere Pro to edit the two original recordings and combine them into a signal video for each treatment group, each approximately 29 minutes long.

To investigate what characteristics of role models enhance or reduce their impacts on students, it is important to ensure that all other aspects of role model sharing remain constant. To achieve this, I design all the treatments to feature the same interviews with the same two role models while selectively and strategically disclosing their characteristics to the

[16]In 2023, about 0.97% of Grade 9 students in the city were admitted by the best local senior high school.
[17]In 2023, around 0.128% of Grade 12 students in the province were admitted to one of these universities.
[18]Similar to Zoom, Tencent Meeting is a software used for online meetings and communications.

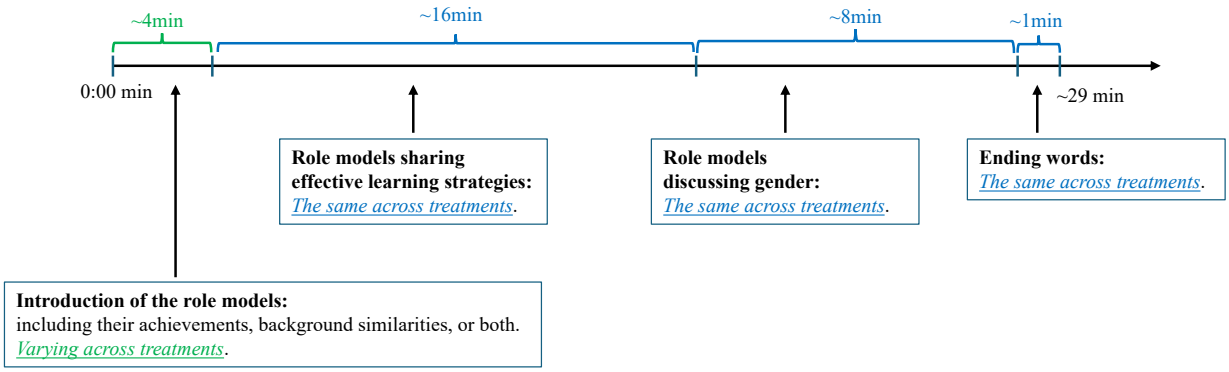


Figure 1: Video timeline (The *green* part differs across the treatments).

students. As summarized in Figure 1, each treatment video spends approximately 4 minutes introducing the two role models to students, which varies across the three treatment arms. After that, the role models shared their effective learning strategies ($\sim$ 16 minutes), discussed gender ($\sim$ 8 minutes), and concluded with one sentence they most wanted to share with the students ($\sim$ 1 minute).

Except for the introduction part, all the role model treatment videos are identical, with each question or topic displayed on the screen and a narration reading it aloud (see Figure 2). Then, the female and male role models respond to the displayed question (see Figure 3 and Figure 4) without a predetermined order of who speaks first. Details about the discussions on effectively learning strategies and gender by the role models during the interview are described in Appendix A.

3.6 Interventions

The interventions of this study consist of three treatment arms and one control arm.

Very Successful & Similar-Background (Very Successful) role models: Students in this treatment arm were assigned to watch a recorded interview featuring the two role models, who are introduced as *very successful* and *having similar backgrounds to the students*. At the beginning, the video fully reveals the achievements of the role models, especially those during their secondary education, and highlights the *background similarities* between role

Figure 2: Each topic or question is displayed on the screen before the role models respond.

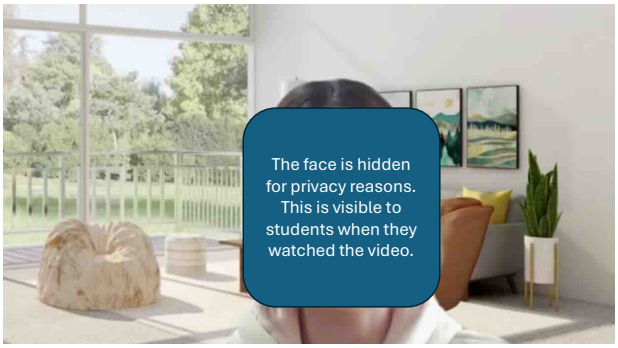


Figure 3: Role models share their experiences and thoughts via online meeting software.

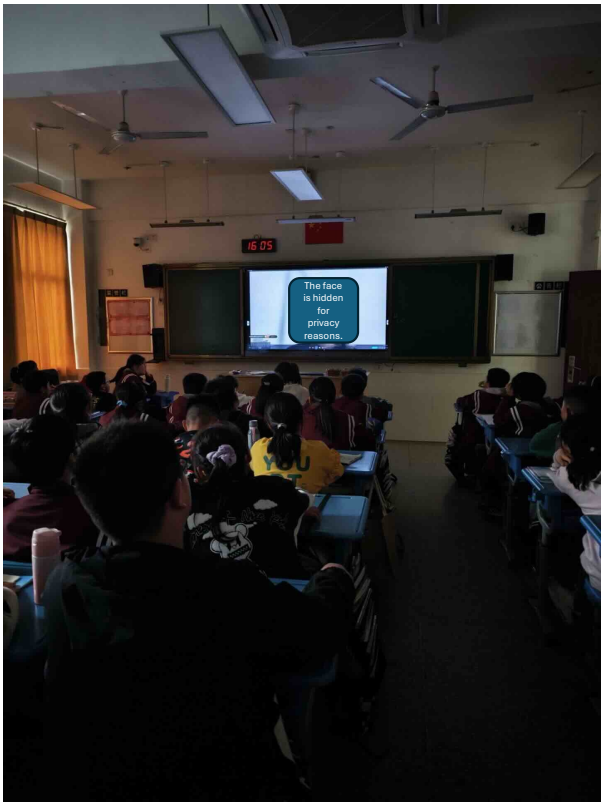


Figure 4: Students were watching the treatment video in the classroom.

models and the students. Following the introduction, the role models share their effective learning strategies and discuss gender as summarized in Section 3.5.

Very Successful & General-Background (VS-General) role models: The recorded interview used in this treatment arm features the same role models as in the Very Successful role model treatment, introducing them as *very successful* by completely revealing their achievements in the same manner as in the Very Successful role model treatment. Yet, it does not explicitly mention the background similarities between the role models and the students. After the introduction, the role models in the VS role model treatment share their effective learning strategies and discuss gender, identical to the content in the Very Successful role model treatment.

Moderately Successful & Similar-Background (Moderately Successful) role models: This treatment arm shows students a recorded interview featuring the same role models

as in the Very Successful and the VS role model treatments. In the video used in this treatment arm, the introduction part highlights the *similar backgrounds shared by the role models and the students*, as in the Very Successful role model treatment. However, I introduce the role models as *moderately successful* by describing their achievements less precisely[19]. The rest of the video, where the role models discuss learning strategies and gender, is identical to the content in the previous two treatment arms.

Control: Students in the control group were assigned to have a non-academic class meeting in the same week as when the treatments took place. The homeroom teachers of controlled classes announced arrangements for the upcoming China Labor Holiday and arranged for students to learn and discuss ways to maintain health and safety when traveling in crowded places during the post-pandemic era.

3.7 Randomization

The treatments of this experiment were randomly assigned. The middle schools received the three treatment videos labeled as "video 1", "video 2", and "video 3", along with an Excel file specifying which video, or no video at all, each class would watch.

The interventions were completed within one week. At the very beginning of that week, the middle schools sent the designated videos directly to the homeroom teachers of the treated classes following the Excel file – these homeroom teachers received only the video assigned to their class and were instructed to play it during the "weekly class meeting" lecture during that week. Homeroom teachers of the controlled classes were instructed to use that same week's weekly class meeting to announce arrangements for the upcoming Chinese Labor Day holiday, along with discussing health and safety measures for traveling in crowded places during the post-pandemic era.

The students were not aware of whether or not they would see a video, nor which video

[19]Instead of stating the fact that these role models ranked top 2 during middle school, the video says that they were among the top 50 students; instead of stating the fact that they attended the best senior high school in the city and the top 2 universities in mainland China, the video says they attended a top 3 high school in the city and a top 10 university in the mainland China.

they would see till the weekly meeting lecture. Each classroom in the recruited middle schools has a computer connected to a big screen in the front, which can play videos with sound. Thus, the treated students watched the video within the classroom to prevent potential interactions between the treatment and control groups.

3.8 Timeline of the Experiment

The experiment was conducted during the Spring 2023 school term, with the timeline presented in Table 1.

TABLE 1 Timeline for the Field Experiment

2023.02.27	***Spring 2023 semester started***.
2023.03-04	***Quality exam***.
2023.04	***Baseline surveys*** to students, their parents, and homeroom teachers.
2023.04	***The interventions*** (three treatment arms and one control arm).
2023.05	***Midterm exam***.
2023.06	***Follow-up survey*** to students.
2023.06-07	***Final exam***.
2023.07	***Spring 2023 semester completed***.

4 Data and Empirical Strategy

4.1 Data

I combine the following sources of data for analysis: (1) students' exam test scores provided by the schools; (2) baseline surveys to students, at least one parent of each student, and homeroom teachers of relevant classes separately; and (3) a follow-up survey to the students.

Test scores: Data on three exams was collected from the middle schools, including one baseline exam before the interventions and two exams after interventions. The baseline exam

is the quality exam, which occurred between late March and early April 2023 – around one month after the Spring 2023 school term started. The baseline test scores are standardized by subtracting the mean and dividing by the standard deviation.

The first post-intervention exam is the midterm exam, which took place in May 2023 – less than one month after the interventions. The second post-intervention exam is the final exam of that semester, which happened in late June 2023 – more than two months after the interventions. Students from each school were tested on the same subjects in each exam, and I dropped the observations if the students did not complete the exam. The post-intervention exams are standardized by subtracting the control mean and dividing by the standard deviation of the students in the control group.[20]

Baseline surveys: The baseline surveys were conducted in April 2023 and were tailored separately for students, their parents, and relevant homeroom teachers. Students were asked self-reported questions primarily about academics, along with a few questions on demographic information. The definitions of key variables used in this paper are detailed in Appendix Table D2. The baseline survey for parents encouraged at least one parent, either the father or the mother, to complete it. This paper does not directly analyze the information provided by the parents but uses their responses to fill in missing baseline demographic data for students. In addition, I include homeroom teacher characteristics as control variables, as their involvement might potentially influence the treatment effects.

Follow-up survey: Students participated in a follow-up survey in June 2023, which was around 2 months after the interventions. I conducted the follow-up survey before the final exam to ensure participation and thus reduce attrition. The follow-up survey covered topics similar to those in the student baseline survey.

This study achieves a participation rate of 96.5%, with 1,920 out of the 1,990 students engaged in the research by having at least one baseline survey completed, either by the students themselves or their parents. These participating students are distributed among 49

[20]The total test score for each exam is calculated by summing up standardized test scores across all subjects and re-normalizing it.

classes across the 5 middle schools (see Appendix Table D1 for details).

4.2 Balance Test

The randomization of this experiment is confirmed by a balance test (see Appendix Table D4). Joint orthogonality tests cannot reject that all the student characteristics are jointly zero for the Very Successful role model treatment (p-value 0.339), the VS-General role model treatment (p-value 0.387), the Moderately Successful role model treatment (p-value 0.876), or any treatment (p-value 0.673). Yet, homeroom teacher characteristics, including their gender and the subjects they teach, are not balanced.[21] The following empirical analysis will control for these homeroom teacher variables.

In my sample, students' beliefs about gender and math ability reflect a biased stereotype, with approximately 58.6% of surveyed students believing that *"Boys are inherently better at math than girls"* in the baseline. However, I do not find significant differences between girls and boys in the pre-intervention math or aggregate test scores. These findings align with evidence from the 2013-2014 China Education Panel Survey (CEPS), a nationally representative survey of Chinese middle school students. The CEPS data shows that around 55.2% of surveyed students agreed that *"Boys are better at math than girls"*, despite girls achieving significantly higher math and total test scores than boys (see Appendix Table D3).

4.3 Attrition

I look at two types of attrition. The first is survey-based attrition, defined as students who did not submit the follow-up survey but had at least one baseline survey completed by either themselves or their parents. The overall attrition rate is 8.5%, with 48 students in the control group, 46 students in the Very Successful role model treatment group, 39 students in the VS-General role model treatment group, and 30 students in the Moderately Successful

[21] The p-values derived from the joint orthogonality tests dropped to 0.074, 0.083, 0.495, and 0.484, respectively, if adding the baseline characteristics of the homeroom teachers.

role model treatment group. Attrition was not differential between the control group and any of the treatment groups, nor among any pairs of the treatment groups (see Column 1 of Appendix Table D5).[22]

I also consider exam-based attrition, defined as students who did not complete the follow-up survey but participated in at least one exam during the Spring 2023 school term. I received students' test scores directly from the schools. If students had taken at least one exam during the Spring 2023 semester, I could get their student IDs and available test scores. Under this measurement, the attrition rate is 11.7%. Similarly, attrition was still balanced among any pairs of the control and treatment groups (see Column 2 of Appendix Table D5).

4.4 Empirical Strategy

To examine the treatment effects, I apply the following regression with school fixed effects:

$$\begin{aligned} Y_{icgs} =& \alpha + \beta_1 \times \textbf{VS} + \beta_2 \times \textbf{VS-General} + \beta_3 \times \textbf{MS} \\ &+ \delta Y_{icgs}^{base} + \mathbf{x'_{icgs}} \cdot \gamma + \mu_s + \varepsilon_{icgs}, \end{aligned} \tag{5}$$

where Y_{icgs} is the outcome of interest for student i from Class c Grade g at School s ($g \in \{7, 8\}$ and $s \in \{1, 2, 3, 4, 5\}$); **VS**, **VS-General**, and **MS** are three dummy variables for treatment assignments[23]; Y_{icgs}^{base} denotes the baseline value of the outcome if measured or excluded otherwise; $\mathbf{x'_{icgs}}$ denotes a set of control variables, including a dummy variable for the student being in Grade 8[24], the gender of class homeroom teachers and whether the homeroom teacher teaches a science subject; μ_s is the school fixed effects; and ε_{igs} indicates that errors

[22]From Appendix Table D6, students with lower baseline test scores or with a homeroom teacher teaching a science subject are less likely to take the follow-up survey. Students with a sibling studying in Grade 7 or 8 at the same middle school are more likely to take the follow-up survey. A joint orthogonality test for attrition cannot reject that all the characteristics are jointly zero when considering student baseline characteristics and their treatment assignment (p-value 0.169).

[23]**VS** stands for Very Successful, and **MS** stands for Moderately Successful.

[24]This was stated as the "grade fixed effects" in the pre-analysis plan. My sample has two grades (Grades 7 and 8), so adding a *Grade8* dummy variable or including grade FE is equivalent statistically. I describe it as a control variable here to avoid potential confusion about FE due to randomization and FE due to the structure of my sample.

are clustered at the unit of randomization, which is the classroom level.

The VS, VS-General, and MS role model treatment effects are estimated by β_1, β_2, and β_3 respectively. Additionally, I investigate the effects of being exposed to more successful role models by testing $(\beta_1 - \beta_3)$, and I examine the effects of revealing similar backgrounds shared between role models and students by testing $(\beta_1 - \beta_2)$.

To correct for multiple hypothesis testing, I calculate the false discovery rate adjusted p-values (*i.e.*, sharpened q-values) of treatment effects β_1, β_2, and β_3. I report the sharpened q-values following the method of Benjamini et al. (2006).

5 Results

5.1 Impacts on Test Scores and Mental Heath

I present the treatment effects on academic performance and mental health issues in Table 2. The Very Successful role models significantly improve academic performance, increasing the midterm total test score by 0.073 standard deviations. This effect becomes stronger in the final exam, conducted around 2 months later, resulting in an increase of 0.085 standard deviations in the final total test score. The Very Successful role models do not significantly affect students' midterm math test score, measured within one month after the intervention, but significantly boost their performance in the final math exam, leading to an increase of 0.184 standard deviations. This suggests that certain skills need time to develop. In addition, the improvement in total test scores, after exposure to the Very Successful role models, is not simply driven by enhanced math performance, as the aggregate test scores excluding math also significantly increase (see Appendix Table D10).

The VS-General role models and the Moderately Successful role models do not have significant impacts on test scores. I plot the cumulative distribution functions (CDFs) of the math and total test scores by treatment or control group (see Appendix Figure D1). The Very Successful role model treatment group shows an observed shift to the right in

the CDF for *all* the test scores. These shifts are observed to be more apparent in the final math test score than in the midterm math test score. I also formally check the equality of the distributions by performing two-sample Kolmogorov-Smirnov tests. Test results reject the equality of the test score distributions when comparing the Very Successful role model treatment group to the control group and the other treatment groups.[25]

The differences in effects between the Very Successful role models and the other two treatment arms are both positive and significant (see Table 2), meaning that two role model characteristics – being more successful and sharing similar backgrounds with the students[26] – enhance their impacts on test scores. Also, the gains from exposure to the more successful role models are more pronounced than those from disclosing the background information.

In Table 2, the poor mental health index is composed of a depression dummy variable and a stress dummy variable, following Anderson (2008). Notably, the control group has 41.6% of depressed students and 34.9% of students under heavy stress in the past two weeks. I find that the Moderately Successful role models effectively reduce the probability of experiencing mental health issues. Compared to the control group, the Moderately Successful role models reduce the likelihood of feeling depressed by 12.0 percentage points, representing a 28.8% decrease in the probability of depression. Similarly, the Moderately Successful role models reduce the probability of being stressed by 9.3 percentage points, representing a 26.6% decrease. I do not observe any significant impacts of the other treatment arms on the mental health variables.

I also compare the impacts of the Very Successful and the Moderately Successful role models, with the results presented as the statistics ($T1-T3$) in Table 2. The results indicate

[25]Test results reject the equality of the test score distributions when comparing the Very Successful role model treatment group to the control group ($pval = 0.056$ for the midterm math, $pval = 0.001$ for the final math, $pval = 0.025$ for midterm total, $pval = 0.005$ for final total), to the VS-General role model treatment group ($pval = 0.012,\ 0.025,\ 0.011,$ and 0.290, respectively), and to the Moderately Successful role model treatment group ($pval = 0.000,\ 0.003,\ 0.004,$ and 0.013, respectively).

[26]Literature has found that similarities in gender, race, and ethnicity enhance the impacts of role models (Kofoed et al., 2019; Beaman et al., 2012) or the impacts of mentorship (Dee, 2004; Fairlie et al., 2014; Eble and Hu, 2020; Gershenson et al., 2022). Nguyen (2008) finds that students are affected by role models with similar socioeconomic status to them. Likewise, I find that similar growth environments shared by role models and role aspirants enhance the impacts of role models.

Table 2: Treatment Effects on the Test Scores and Mental Health

	(1) Midterm Math	(2) Final Math	(3) Midterm Total	(4) Final Total	(5) Poor Mental Health Index	(6) Depression Dummy	(7) Stress Dummy
T1: Very Successful	0.076	0.184**	0.073***	0.085***	0.004	0.003	0.000
	(0.076)	(0.069)	(0.027)	(0.030)	(0.115)	(0.051)	(0.051)
	[0.977]	[0.055]*	[0.049]**	[0.038]**	[0.999]	[0.999]	[0.999]
T2: VS-General	-0.045	0.021	0.014	0.044	-0.013	-0.022	0.010
	(0.071)	(0.060)	(0.026)	(0.034)	(0.098)	(0.046)	(0.042)
	[0.977]	[0.999]	[0.999]	[0.559]	[0.999]	[0.999]	[0.999]
T3: Moderately Successful	-0.066	0.001	-0.018	-0.011	-0.252***	-0.120***	-0.093**
	(0.074)	(0.062)	(0.029)	(0.038)	(0.077)	(0.033)	(0.037)
	[0.977]	[0.999]	[0.999]	[0.999]	[0.010]***	[0.004]***	[0.089]*
Observations	1,857	1,835	1,857	1,835	1,726	1,726	1,726
R-squared	0.742	0.636	0.862	0.771	0.022	0.024	0.018
Effects of revealing different role model characteristics:							
BG=*T1-T2*	0.121	0.163**	0.059*	0.041	0.017	0.025	-0.010
	(0.088)	(0.072)	(0.032)	(0.030)	(0.120)	(0.055)	(0.050)
More successful=*T1-T3*	0.142	0.183**	0.091***	0.096***	0.256***	0.123**	0.093*
	(0.093)	(0.071)	(0.033)	(0.034)	(0.111)	(0.048)	(0.051)
control mean						0.416	0.349

"Midterm Total" or "Final Total" refers to the standardized aggregate score across all subjects taken in the midterm or final exam. Standardized aggregate scores are composed of subject-standardized scores and normalized. "Poor Mental Heath Index" is a standardized weighted average of "Depression Dummy" and "Stress Dummy", following Anderson (2008). Any missing baseline score is replaced by the median pre-intervention exam score, and a dummy variable is included to capture this. Control variables include student baseline test score, a dummy variable for the student performing below the median within grade 7/8 at her school in the baseline exam, homeroom teacher gender, a dummy variable for the homeroom teacher teaching a science subject, and a dummy variable indicating the student is in Grade 8. Columns (5)-(7) also control for a dummy variable for the homeroom teacher teaching Chinese literature. All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). Sharpened q-value in square brackets, following the method of Benjamini et al. (2006). *** p<0.01, ** p<0.05, * p<0.1.

that role models with varying success levels influence students in significantly different ways. While higher-achieving role models enhance students' academic performance, the moderately successful role models improve their mental health.

5.1.1 Robustness checks

To confirm the robustness of my findings, I perform the following additional tests.

Permutation test I conducted a permutation test with 10,000 repetitions to address the concerns about a relatively small number of clusters. I find that the permutation p-values reject the null hypotheses at the same levels as the robust p-values (see Appendix Table D8).

Alternative specifications I show that the results are robust to alternative specifications, including regressions without the Grade8 dummy variable (*i.e.,* without grade fixed effects), regressions that control for baseline value of outcome and the Grade8 dummy variable, and regression that control for baseline value of outcome only (see Appendix Table D9).[27]

Scope for Spillovers Students in the control group may have learned about the role models from their siblings who study in a treated class or from their friends in a treatment group. If the direction of the treatment effects and that of the potential spillover effects are the same, the existence of the spillovers would underestimate the overall role model treatment effects identified in this study. I estimate the scope for spillovers in this experiment and find the scope is quite limited (see Appendix Table D7). Only 3.9% of the students in a controlled class have siblings in Grade 7 or 8 at their school, and 3.1% of the controlled students reported that their siblings are in the same grade as them[28]. Students might mention the role models when talking to friends in other classes. I find that, despite many of them having discussed aspirations with family members and classmates[29], only 14.7% of the students in the control

[27]I also perform a logit regression to check the robustness of the treatment effects on the depression and stress dummy variables. The results of logit regressions confirm the robustness of my findings (see Appendix Table D11).

[28]I might overestimate the scope of the spillovers through the sibling channel, as siblings in the same grade are typically assigned to the same class. According to the middle schools, this practice helps prevent scheduling conflicts for parents, such as the need to attend parent-teacher meetings.

[29]This result in my experiment reflects the previous finding that most peer interactions happen within class (Avvisati et al., 2014).

group have discussed their aspirations and future goals with their friends from other classes.

5.1.2 Impacts on long-term educational outcomes: Pursuing further education

Upon completing Grade 9, middle school students can take the Senior High School Entrance Exam (also known as *Zhong-Kao*), where their performance will determine their eligibility for senior high school education and the types of senior high schools to which they can apply. Analyzing these long-term test outcomes reveals whether the role model treatments influence students' ambitions and their ability to continue their education.

In the sub-sample of Grade 8 students[30], the Senior High Entrance Exam take-up rate reaches 98%. Receiving any role model treatment raises the likelihood of taking the Entrance Exam by 2.8 percentage points.

Students in the Very Successful role model treatment group experienced a significant increase of 0.189 standard deviations in their overall test score in the Senior High School Entrance Exam (see Appendix Table C1). This improvement in academic performance is observed across all subjects taken, with the math test scores increasing by 0.338 standard deviations and the total scores excluding math rising by 0.182 standard deviations. Consequently, these enhanced long-term test scores translate into a 9.7 percentage point increase in the probability of eligibility for applying to vocational schools.

The VS-General role models significantly improve the total test score by boosting performance in subjects other than math, although this impact is less pronounced than that of the Very Successful role models. Consistent with their effects in the short term, the Moderately Successful role models do not have significant impacts on the results of the Senior High School Entrance Exam.[31]

[30]Grade 7 students will complete their middle school study and choose whether to take the Senior High School Entrance Exam in 2025. Results will be updated after I receive the long-term educational results of these Grade 7 students.

[31]Details about the role model impacts on long-term educational outcomes, along with multiple relevant robustness checks, are discussed in Appendix C.

5.2 Treatment Effects by Gender

Previous discussions reveal a trade-off due to exposure to role models of different success levels, as exposure to higher-achieving role models enhanced their academic performance, whereas exposure to moderately-achieving role models improved their mental health. This trade-off between academic performance enhancement and mental health improvement is found among girls and boys (see Table 3).

From Panel A in Table 3, I find that the Very Successful role models improve girls' math test scores by 0.115-0.225 standard deviations and their overall test scores by 0.090-0.094 standard deviations. I do not find any significant changes in girls' test scores in the other treatment groups. Compared to the other treatments, the Very Successful role models significantly improve girls' exam outcomes, indicating that higher-achieving role models have a more pronounced effect on enhancing the academic performance of girls.

In my sample, approximately 51.1% of girls in the control group reported recently experiencing depression, and 35.5% reported stress. These proportions are higher compared to those of boys in the control group. The Moderately Successful role models significantly reduce the likelihood of having mental health issues among girls, decreasing the probability of depression by 32.1% and that of stress by 19.0%.[32] I compare the effects of the Very Successful and the Moderately Successful role models on girls' mental health, *i.e.*, the statistics $(T1 - T3)$, and find that the difference in these treatments are significantly large. This indicates that the Very Successful role models influence girls' mental health in a very different way compared to the mental health relief provided by the Moderately Successful role models.

Boys also experienced this trade-off, while their educational outcome enhancement from the higher-achieving role models is smaller than girls (see Panel B in Table 3). The Very Successful role models improve their math test scores by 0.037-0.143 standard deviations and their overall test scores by 0.058-0.078 standard deviations. The Moderately Successful role

[32] I check the robustness of these findings using logit regressions. Results are presented in Appendix Table D11, which do not reject the findings presented here.

Table 3: Treatment Effects on the Test Scores and Mental Health by Gender

	(1) Midterm Math	(2) Final Math	(3) Midterm Total	(4) Final Total	(5) Poor Mental Health Index	(6) Depression Dummy	(7) Stress Dummy
Panel A. Girls							
T1: Very Successful	0.115	0.225***	0.090***	0.094**	0.038	0.010	0.022
	(0.076)	(0.072)	(0.030)	(0.037)	(0.140)	(0.064)	(0.061)
	[0.739]	[0.016]**	[0.026]**	[0.071]*	[0.999]	[0.999]	[0.999]
T2: VS-General	-0.032	0.031	0.031	0.039	0.023	-0.062	0.080
	(0.067)	(0.065)	(0.029)	(0.046)	(0.142)	(0.064)	(0.062)
	[0.999]	[0.999]	[0.790]	[0.999]	[0.999]	[0.937]	[0.559]
T3: Moderately Successful	-0.053	-0.019	0.009	0.022	-0.273**	-0.164***	-0.067
	(0.075)	(0.083)	(0.037)	(0.056)	(0.107)	(0.048)	(0.050)
	[0.999]	[0.999]	[0.999]	[0.999]	[0.075]*	[0.008]***	[0.559]
Observations	907	904	907	904	836	836	836
R-squared	0.748	0.660	0.862	0.762	0.040	0.065	0.020
control mean	-0.133	-0.133	-0.065	-0.084	0.116	0.511	0.353
Effects of revealing different role model characteristics:							
BG=*T1-T2*	0.147*	0.194***	0.059*	0.055	0.015	0.071	-0.057
	(0.077)	(0.070)	(0.033)	(0.040)	(0.136)	(0.062)	(0.059)
More successful=*T1-T3*	0.168*	0.244***	0.081**	0.072	0.311***	0.173***	0.090*
	(0.087)	(0.085)	(0.039)	(0.051)	(0.109)	(0.048)	(0.053)
Panel B. Boys							
T1: Very Successful	0.037	0.143*	0.058*	0.078**	-0.038	-0.011	-0.021
	(0.080)	(0.076)	(0.033)	(0.038)	(0.117)	(0.053)	(0.054)
	[0.999]	[0.375]	[0.470]	[0.260]	[0.999]	[0.999]	[0.999]
T2: VS-General	-0.062	-0.000	-0.002	0.046	-0.061	0.010	-0.061
	(0.081)	(0.069)	(0.036)	(0.041)	(0.084)	(0.041)	(0.039)
	[0.999]	[0.999]	[0.999]	[0.627]	[0.999]	[0.999]	[0.341]
T3: Moderately Successful	-0.081	0.018	-0.045	-0.046	-0.234**	-0.077*	-0.120**
	(0.081)	(0.065)	(0.032)	(0.048)	(0.093)	(0.040)	(0.045)
	[0.999]	[0.999]	[0.470]	[0.627]	[0.085]*	[0.331]	[0.055]*
Observations	949	930	949	930	890	890	890
R-squared	0.740	0.617	0.864	0.783	0.021	0.011	0.030
control mean	0.133	0.127	0.090	0.092	-0.107	0.329	0.346
Effects of revealing different role model characteristics:							
BG=*T1-T2*	0.099	0.143	0.061	0.032	0.023	-0.021	0.040
	(0.103)	(0.089)	(0.039)	(0.039)	(0.122)	(0.060)	(0.051)
More successful=*T1-T3*	0.118	0.124	0.103***	0.124***	0.196	0.066	0.099
	(0.105)	(0.085)	(0.034)	(0.045)	(0.132)	(0.059)	(0.059)

Panel A shows the results for girls and Panel B presents the results for boys. "Midterm Total" or "Final Total" refers to the standardized aggregate score across all subjects taken in the midterm or final exam. Standardized aggregate scores are composed of subject-standardized scores and normalized. "Poor Mental Heath Index" is a standardized weighted average of "Depression Dummy" and "Stress Dummy", following Anderson (2008). "Baseline Math" or "Baseline Total" refers to the standardized score achieved in the pre-intervention exam. Any missing baseline score is replaced by the median pre-intervention exam score, and a dummy variable is included to capture this. Control variables include student baseline test score, a dummy variable for the student performing below the median within grade 7/8 at her school in the baseline exam, homeroom teacher gender, a dummy variable for the homeroom teacher teaching a science subject, and a dummy variable indicating the student is in Grade 8. Columns (5)-(7) also control for a dummy variable for the homeroom teacher teaching Chinese literature. All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). Sharpened q-value in square brackets, following the method of Benjamini et al. (2006). *** p<0.01, ** p<0.05, * p<0.1.

models also significantly decrease the probability of having mental health issues among boys. Unlike the impacts on girls, the Very Successful role models slightly reduce the likelihood of having mental health issues among boys. I do not find significant differences between the impacts of the Very Successful and the Moderately Successful role models on boys' mental health, meaning that higher-achieving role models do not impose additional mental health burdens on boys compared to moderately successful role models.

5.3 Treatment Effects by Baseline Ability

The impacts of the Very Successful role models on girls' mental health is significantly different from the mental health relief provided by the Moderately Successful role models. To investigate, I divide the students into four groups based on their baseline total test score: Q1 (bottom-performing), Q2 (lower-middle performing), Q3 (upper-middle performing), and Q4 (top-performing). I break down the treatment effects by interacting the treatment dummy variables with each quartile of the baseline score.

The bottom-performing girls experienced trade-off effects due to exposure to higher-achieving role models (see Appendix Table D13). These bottom-performing girls, after being exposed to the Very Successful role models, significantly improved their total test score by 0.200 standard deviations in the midterm exam but also experienced a significant increase in their poor mental health index by 0.385 standard deviations. The positive impacts on test scores diminished in the final exam, with an improvement of 0.149 standard deviations among the bottom-performing girls. Compared to the Moderately Successful role models, the Very Successful role models improve the bottom-performing girls' midterm overall exam score by 0.149 standard deviations, their final total score by 0.077 standard deviations, but also raise their poor mental health index by 0.633 standard deviations. These findings indicate that these bottom-performing girls benefit from the higher-achieving role models in improving their academic performance, but it comes with a cost to their mental health.[33]

[33]I also check the heterogeneity in treatment effects by baseline ability among the boys. Results are shown

Higher-achieving role models enhance the math exam performance among girls above the median, with significant increases of 0.261 standard deviations for upper-middle performing girls and 0.213 standard deviations for top-performing girls in their final math exam. These improvements, driven by the Very Successful role models, significantly exceed those influenced by the Moderately Successful role models. Compared to the Moderately Successful role models, the Very Successful role models increase the final math exam by 0.319 standard deviations for upper-middle performing girls and by 0.249 standard deviations for top-performing girls. However, I do not observe any changes in their mental health, suggesting that the relatively high-performing girls do not experience any trade-off effects due to exposure to the Very Successful role models.

I use one machine learning approach named *causal forest*, developed by Wager and Athey (2018), to identify the baseline characteristics of students who responded most from the Very Successful role model treatment (see Appendix B for details). Dividing the treatment heterogeneity into four quartiles, I compare the average baseline characteristics between those in the top quartile and those in the bottom quartile. Results from the *causal forest* approach confirm my previous findings that the bottom-performing girls experienced the trade-off effects after receiving the Very Successful role model treatment (see Appendix Table D12). Moreover, my findings from the causal forest analysis indicate that students who experienced the greatest test score gains but the most significant declines in mental health after receiving the Very Successful role model treatment had lower baseline test scores and spent fewer hours studying at baseline. These students are more likely to be girls, have siblings, and have a male homeroom teacher.[34]

in Appendix Table D15. I do not find the existence of the trade-off effects in any subgroups of boys.

[34]This finding aligns with the literature indicating that students benefit from being assigned a homeroom teacher who shares similar identities with them, such as gender or ethnicity (Dee, 2004; Fairlie et al., 2014; Gershenson et al., 2022) and that assigning female math teachers benefits girls with low perceived abilities (Eble and Hu, 2020).

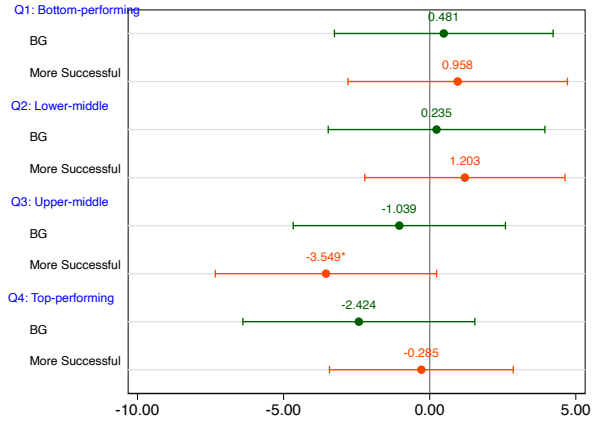
5.4 Why Do Girls Improve Test Scores but Still Struggle?

After exposure to the Very Successful role models, the bottom-performing girls observed their improved test scores in the midterm exam but still felt depressed and stressed. To understand why, I examine possible mechanisms, including their learning effort, aspirations, and beliefs. The results are shown in Appendix Table D14, with the effects of exposure to different characteristics of role models being visualized in Figure 5.

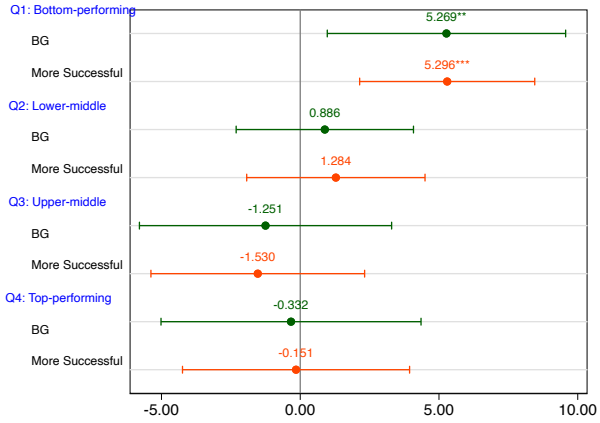
The Very Successful role models encourage the bottom-performing girls to spend an additional 3.290 hours per week studying. In contrast, the Moderately Successful role models have little impact on the learning effort. The differences between these two role model treatments are significantly large, with a gap of 5.296 hours per week (see Figure 5-b). This increased learning effort, driven by the Very Successful role models, primarily results from more weekday study hours. However, the increase in overall study effort is not due to more time spent studying math (see Figure 5-a) or changes in beliefs about gender and learning ability (see Figure 5-d).

The Very Successful role models also encourage bottom-performing girls to adopt higher aspirations for further education. They increase the number of bottom-performing girls who aspire to rank above the median in class in the Senior High School Entrance Examination, and such increase is significant compared to the control group and the other two treatment groups (see Figure 5-c). Nevertheless, despite their increased test scores, these bottom-performing girls did not find any significant improvement in their rankings in the midterm exam (see Appendix Table D16), suggesting that these girls also realized the elevated goals were hard to reach. This might be because their classmates were also exposed to the same role model treatment or because the gaps between the bottom-performing girls and others were large prior interventions.

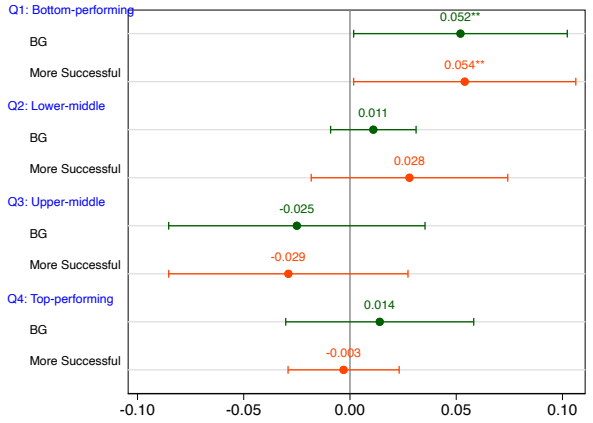
Figure 5: Role Model Effects on Girls' Efforts, Aspirations, and Beliefs by Baseline Exam Quartiles



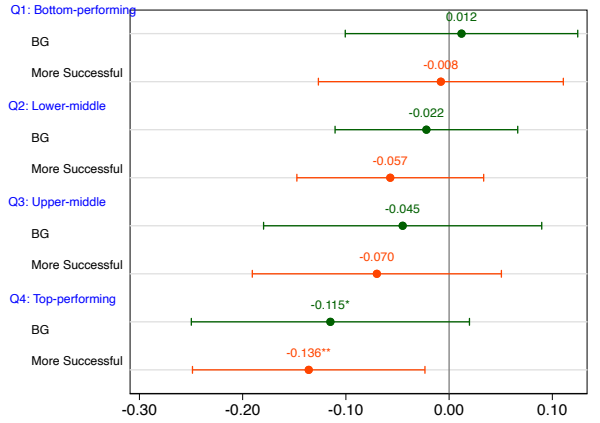
(a) Effort on studying math.



(b) Effort on studying all subjects.



(c) Aspire to rank above median in the Senior High School Entrance Examination.



(d) Beliefs about whether boys are inherently better at math than girls.

Quartile refers to the quartile of the pre-intervention exam scores, from Q1 (bottom-performing) to Q4 (top-performing). "Effort on studying math" or "Effort on studying all subjects" refers to the total hours spent during a week on studying math or all subjects. Aspiration is defined as a dummy for "Aspire to rank above median in the Senior High School Entrance Examination". "Belief" is a dummy variable for the student who believes that boys are inherently better at math than girls. Control variables include student gender, student baseline aggregate test score, student baseline response to each question if it exists, homeroom teacher gender, a dummy variable for the homeroom teacher teaching a science subject, and a dummy variable indicating the student is in Grade 8. Regressions on overall effort and aspiration also control for a dummy for the homeroom teacher teaching Chinese literature. Regressions on efforts spent on studying math and beliefs also control for student pre-intervention math test scores. All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1. The regression table is shown in Appendix Table D14.

6 Policy Implications

Addressing Mental Health in Role Model Interventions: Higher-achieving role models are more effective in boosting academic performance. However, my study shows that exposure to more successful role models negatively influences the mental health of an underperforming subgroup of role aspirants. This occurs when students invest more effort but still find their improved educational outcomes falling short of their elevated aspirations. This paper highlights the need to consider mental health when implementing and evaluating role model interventions. To support students' mental health, schools and policymakers should provide accessible mental health resources, such as stress management workshops and counseling services.

Implications for educational attainment enhancement in China: Despite rapid economic growth, educational attainment in China's labor force remains notably low compared to countries with lower GDPs (Khor et al., 2016). Poor academic performance is often associated with an increased likelihood of school dropout (Shi et al., 2015) and poor mental health (Kötter et al., 2017). In addition, a large proportion of students and their parents continue to hold the belief that boys are better at math than girls, even though girls no longer consistently fall behind boys in science subjects (see Appendix Table D3). To address these issues and improve educational outcomes, my paper suggests that schools could invite high-achieving role models to share their learning strategies and challenge inaccurate views about gender and learning abilities.

Around 74% of rural students in China are at risk for mental health issues, which are correlated with a high likelihood of school dropout (Wang et al., 2015). To address this, my paper suggests inviting moderately achieving role models to improve and support students' mental health. [35]

Cost-effectiveness: This study is remarkably cost-effective, with the interventions imple-

[35]Due to the randomized school enrollment process (details discussed in Section 3.2), the findings and implications of this study can be applicable on a larger scale beyond just the five middle schools involved.

mented at no cost. The role models were completely voluntary and received no compensation for being interviewed.[36] Students in the treatment groups watched the videos in their classrooms, which are equipped for video playing, making the intervention convenient and easy for schools to implement. This study presents a significant advancement over previous low-cost interventions aimed at improving students' academic performance (Banerjee et al., 2007; Riley, 2019) and is comparable to the statistics intervention used in (Nguyen, 2008). Scaling up these role model interventions by inviting their successful alumni can enhance the cost-effectiveness of this approach.

7 Conclusion

Exposure to role models with varying success levels leads to a trade-off, as higher-achieving role models enhance students' academic performance, whereas moderately achieving role models improve their mental health. Students exposed to the Very Successful role models improved test scores by 0.07-0.18 standard deviations, whereas those exposed to the Moderately Successful role models experienced a reduced probability of feeling depressed and stressed by 29.6% and 26.6%, respectively.

Further, I find that the Very Successful role models significantly improve bottom-performing girls' midterm total score by 0.2 standard deviations but increase their poor mental health index by 0.385 standard deviations. These bottom-performing girls, after exposure to the Very Successful role models, invested more effort in learning but still found their improved educational outcomes falling short of their elevated goals.

Leveraging cost-effective treatments, this paper presents the first empirical study to highlight the negative impacts of role models on mental health as a trade-off for enhancing academic performance in underperforming subgroups. It emphasizes the need to consider mental health when implementing role model interventions and support mental health in

[36]To ensure high-quality video editing, I subscribed to Adobe Premiere Pro at a monthly cost of $35. However, in practice, open-source software or no editing also works if role aspirants have sufficient time to watch videos.

underperforming subgroups. Future work could expand beyond the educational context to explore how these findings translate into different environments.

This paper also brings policy implications for improving educational attainment in China. It addresses the low educational attainment rate in China (Khor et al., 2016) by suggesting schools invite high-achieving role models to boost students' educational outcomes. Mental health issues, correlated with high school dropout rates, are prevalent among rural Chinese students (Wang et al., 2015). To tackle this issue, my paper suggests inviting moderately achieving role models to support students' mental health. Scaling up these role model interventions by encouraging schools to invite their successful alumni can further enhance the cost-effectiveness of this approach.

References

Marcos Agurto, M Bazan, S Hari, and S Sarangi. Women in engineering: The role of role models. Technical report, GLO Discussion Paper, 2021.

Michael L Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495, 2008.

Francesco Avvisati, Marc Gurgand, Nina Guyon, and Eric Maurin. Getting parents involved: A field experiment in deprived schools. *Review of Economic Studies*, 81(1):57–83, 2014.

Abhijit V Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden. Remedying education: Evidence from two randomized experiments in india. *The quarterly journal of economics*, 122(3):1235–1264, 2007.

Lori Beaman, Esther Duflo, Rohini Pande, and Petia Topalova. Female leadership raises aspirations and educational attainment for girls: A policy experiment in india. *science*, 335(6068):582–586, 2012.

Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

Tanguy Bernard, Stefan Dercon, Kate Orkin, Alemayehu Taffesse, et al. *The future in mind: Aspirations and forward-looking behaviour in rural Ethiopia*, volume 10224. Centre for Economic Policy Research London, 2014.

Tanguy Bernard, Stefan Dercon, Kate Orkin, and Alemayehu Seyoum Taffesse. Parental aspirations for children's education: Is there a "girl effect"? experimental evidence from rural ethiopia. *AEA Papers and Proceedings*, 109:127–132, 2019.

Tanguy Bernard, Stefan Dercon, S Orkin, Giulio Schinaia, and A Seyoum Taffesse. The future in mind: aspirations and long-term outcomes in rural ethiopia. 2023.

Rebecca A Bernert, Katherine A Merrill, Scott R Braithwaite, Kimberly A Van Orden, and Thomas E Joiner Jr. Family life stress and insomnia symptoms in a prospective evaluation of young adults. *Journal of Family Psychology*, 21(1):58, 2007.

Thomas Breda, Julien Grenet, Marion Monnet, and Clémentine Van Effenterre. How effective are female role models in steering girls towards stem? evidence from french high schools. *The Economic Journal*, 133(653):1773–1809, 2023.

Nilanjana Dasgupta and Shaki Asgari. Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of experimental social psychology*, 40(5):642–658, 2004.

Jonathan MV Davis and Sara B Heller. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–550, 2017.

Thomas S Dee. Teachers, race, and student achievement in a randomized experiment. *Review of economics and statistics*, 86(1):195–210, 2004.

Aurelia Di. The power of words: How role models influence attitudes and behaviors. *Available at SSRN: https://ssrn.com/abstract=4749301 or http://dx.doi.org/10.2139/ssrn.4749301*, 2024.

Alex Eble and Feng Hu. Child beliefs, societal beliefs, and teacher-student identity match. *Economics of Education Review*, 77:101994, 2020.

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

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

Garance Genicot and Debraj Ray. Aspirations and economic behavior. *Annual Review of Economics*, 12(1):715–746, 2020.

Seth Gershenson, Cassandra MD Hart, Joshua Hyman, Constance A Lindsay, and Nicholas W Papageorge. The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy*, 14(4):300–342, 2022.

Jennifer Golan and Jing You. Raising aspirations of boys and girls through role models: Evidence from a field experiment. *The Journal of Development Studies*, 57(6):949–979, 2021.

Robert Jensen. Do labor market opportunities affect young women's work and family decisions? experimental evidence from india. *The Quarterly Journal of Economics*, 127(2):753–792, 2012.

Robert Jensen and Emily Oster. The power of tv: Cable television and women's status in india. *The Quarterly Journal of Economics*, 124(3):1057–1094, 2009.

Niny Khor, Lihua Pang, Chengfang Liu, Fang Chang, Di Mo, Prashant Loyalka, and Scott Rozelle. China's looming human capital crisis: upper secondary educational attainment rates and the middle-income trap. *The China Quarterly*, 228:905–926, 2016.

Elijah Kipkech Kipchumba, Catherine Porter, Danila Serra, Munshi Sulaiman, et al. Infuencing youths' aspirations and gender attitudes through role models: Evidence from somali schools. Technical report, 2021.

Michael S Kofoed et al. The effect of same-gender or same-race role models on occupation choice: evidence from randomly assigned mentors at west point. *Journal of Human Resources*, 54(2): 430–467, 2019.

Thomas Kötter, Josefin Wagner, Linda Brüheim, and Edgar Voltmer. Perceived medical school stress of undergraduate medical students predicts academic performance: an observational study. *BMC medical education*, 17:1–6, 2017.

Jeanne Lafortune, Julio Riutort, and José Tessada. Role models or individual consulting: The impact of personalizing micro-entrepreneurship training. *American Economic Journal: Applied Economics*, 10(4):222–245, 2018.

Patrick Lubega, Frances Nakakawa, Gaia Narciso, Carol Newman, Archileo N Kaaya, Cissy Kityo, and Gaston A Tumuhimbise. Body and mind: Experimental evidence from women living with hiv. *Journal of Development Economics*, 150:102613, 2021.

Travis J Lybbert and Bruce Wydick. Poverty, aspirations, and the economics of hope. *Economic development and cultural change*, 66(4):709–753, 2018.

Thekla Morgenroth, Michelle K Ryan, and Kim Peters. The motivational theory of role modeling: How role models influence role aspirants' goals. *Review of general psychology*, 19(4):465–483, 2015.

Trang Nguyen. Information, role models and perceived returns to education: Experimental evidence from madagascar. 2008.

Emma Riley. Role models in movies: the impact of queen of katwe on students' educational attainment. *The Review of Economics and Statistics*, pages 1–48, 2019.

Emma Riley. Resisting social pressure in the household using mobile money: Experimental evidence on microenterprise investment in uganda. *American Economic Review*, 114(5):1415–1447, 2024.

Danila Serra. Role models in developing countries. *Handbook of experimental development economics*, 2022.

Yaojiang Shi, Linxiu Zhang, Yue Ma, Hongmei Yi, Chengfang Liu, Natalie Johnson, James Chu, Prashant Loyalka, and Scott Rozelle. Dropping out of rural china's secondary schools: A mixed-methods analysis. *The China Quarterly*, 224:1048–1069, 2015.

Jane G Stout, Nilanjana Dasgupta, Matthew Hunsinger, and Melissa A McManus. Steming the tide: using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (stem). *Journal of personality and social psychology*, 100(2):255, 2011.

Matthew A Stults-Kolehmainen and Rajita Sinha. The effects of stress on physical activity and exercise. *Sports medicine*, 44:81–121, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Huan Wang, Chu Yang, Fei He, Yaojiang Shi, Qinghe Qu, Scott Rozelle, and James Chu. Mental health and dropout behavior: A cross-sectional study of junior high students in northwest rural china. *International Journal of Educational Development*, 41:1–12, 2015.

Appendices

A Overview of the Role Model Interviews

The male and female role models in this study were interviewed separately to avoid any interactions between the role models that might affect their responses. Both role models received ***the same*** question list before the interview, in which they were asked to share their effective learning strategies and discuss gender. When they received the list of questions, the role models were informed that their interviews would be shared with local middle school students. However, they were not told which specific middle schools would receive the interview or the exact purpose of this experiment. The interviews with the two role models were conducted using Zoom or Tencent Meeting[37], respectively. At the very beginning of their interview and before they spoke, the role models were informed that their responses would be recorded by the online meeting software in use. The interviewer did not implicitly direct or manipulate the responses of role models. Everyone in the interview spoke Chinese (Pu Tong Hua), which all the middle school students in my experiment can understand and follow. Interviewing the role models gave two original video recordings.

I clipped these two original recordings with Adobe Premiere Pro, a professional video editing software. The Adobe Premiere Pro software requires a subscription fee of $35 per month. This can be substituted with other open-source video editing software, or the editing step can be skipped entirely if students have sufficient time to watch videos. The clipped video for each treatment is approximately 29 minutes long, featuring the contains as detailed in the Interventions Section (or Section 3.6). The format of the clipped role model interviews is as follows: after introducing the role models, each question or topic is displayed on the screen with a narration reading the question aloud (see Figure 2); and then, the female and male role models responded to the question displayed (see Figure 3 and Figure 4). There is no specific order in which role models should speak first. Each clipped treatment video has three main contains as described in the following:

Introducing the role models: *Approximately 4 minutes; Varying across the three treatments.* For the Very Successful and the VS-General role model treatments, the achievements of the role models

[37]Tencent Meeting is a software similar to Zoom or Microsoft Teams, used for online meetings and communications.

are fully disclosed, including ranking Top 2 at middle school, going to the best senior high school in the city, going to a Top 2 university in China, and their current occupations. For the Moderately Successful role model treatment, their achievements are described less precisely, including ranking in the top 50 in middle school, going to a Top 3 senior high school in the city, going to a Top 10 university in China, and their current occupations. In addition, the Very Successful and the Moderately Successful role model treatments also fully disclose the background similarities between role models and the students, where the role models are asked to briefly describe a typical school day during their middle school. In these two treatment groups, the role models described their routines and also identified the landmarks and the street with food and entertainment options near the middle schools. In the VS-General role model treatment, where the information about background similarities is not disclosed, the description of a typical school day is clipped out.

Sharing effective learning strategies: *Around 16 minutes; The same across the three treatments.* In this session, the role models were asked to share effective learning strategies, including how to fully utilize after-class exercises and tutoring classes, how to set ambitious yet feasible aspirations, how to strive to achieve their goals properly, and how to manage study pressure or overcome the frustrating feelings when their performance did not meet expectations. The role models also shared some beneficial habits they developed during secondary education that helped them manage time effectively. In their interview, the role models emphasized the importance of *knowing yourself* rather than *simply copying and pasting methods of others*. *Hard work is appreciated* but it is also crucial to *identify and improve upon one's shortcomings*. Additionally, they encouraged *being confident enough to aim high and challenge oneself with even higher goals*, while also *being persistent enough to make a difference*. However, it is common that *people might not be able to realize every goal. Regret only arises if they did not try.*

Discussing gender: *Approximately 8 minutes; The same across the three treatments.*

In this part, the role models shared their opinions on a few gender-biased statements in education, including "Boys are inherently smarter than girls", "Boys are inherently better at science subjects than girls", and "Boys will improve faster and more in senior grades than girls, even if they do not work hard during the junior years". The role models also discussed how to empower oneself rather than being negatively affected by biased beliefs about gender. Using their own experiences,

both role models disagreed with any of the above statements, saying that *it is about competition among remarkable people, not between genders.* They encouraged *critical thinking* – people have strengths and shortcomings because they are unique, not because of their gender or external opinions. Conversely, *individuals need to determine who they want to be* and *work diligently to achieve that, instead of letting others define their lives or prescribe who they should be.*

B Causal Forest – A Machine Learning Approach

To check the robustness of my heterogeneity analysis results in treatment effects by baseline total test score, I perform a machine learning approach, *casual forest*, as developed in Wager and Athey (2018), implemented in Davis and Heller (2017), and widely applied, including in Riley (2024). This approach also allows me to analyze numerous sources of heterogeneity simultaneously, indicating additional sources of heterogeneity in the effects of role model treatment.

I use the `grf` command from the `mlrtime` package in STATA, which calls the `causal_forest` function from the `grf` package in R. To run the casual forest algorithm, the dataset was randomly split into two subsamples – one subsample used to train the causal forest and the other subsample used to estimate the average treatment effects. Then, I follow Riley (2024) to construct quartiles based on the estimated treatment effects and compare the differences of mean characteristics for the students in the top and bottom quartiles.

C Long-Term Outcomes – Continuation in Education

Upon completing Grade 9, middle school students can take the Senior High School Entrance Exam (also known as *Zhong-Kao*), where their performance will determine their eligibility for upper secondary education and the types of senior high schools to which they can apply. I analyze these long-term test outcomes to learn whether the role model treatments influence students' ambitions and their ability to continue their education.

I am able to follow up with three out of the five middle schools[38] recruited for this experiment

[38]Tracking long-term exam performance was not included in the pre-analysis plan and thus requires additional permissions from each middle school. I got permission from Middle Schools 2, 3, and 4, as listed in

to access students' results in the Senior High School Entrance Exam. The empirical analysis in this section utilizes the Entrance Exam test scores of *Grade 8 students only*. Results will be updated after Grade 7 students participate in the Senior High School Entrance Exam in 2025.

In the sub-sample of Grade 8 students from these three middle schools, 98% of the students participated in the Senior High School Entrance Exam in 2024. In the control group, the take-up rate for the Senior High School Entrance Exam reached 96%. Receiving any role model treatment raises the probability of taking this exam by 2.8 percentage points, indicating a 2.9% increase in the probability of taking the Senior High School Entrance Exam.

I show results for the long-term role model treatment effects on test scores in Columns (1)-(4) of Appendix Table C1. I find that students who were exposed to the Very Successful role models experienced a significant increase of 0.189 standard deviations in their overall test scores gained in the Senior High School Entrance Exam. This improvement is observed as an improvement in all subjects taken, with the math test scores increasing by 0.338 standard deviations and the total scores excluding math rising by 0.182 standard deviations. Similar to the short-term effects on academic performances, the Moderately Successful role models have little influence on the long-term educational outcomes. The statistics $(T1 - T3)$ are significantly large, suggesting that the Very Successful and the Moderately Successful role models affect the students differently.

I also test the long-term impacts of the role model treatments on girls and boys, respectively. Results are presented in Columns (5)-(12) of Appendix Table C1. The impacts on these long-term academic outcomes are consistent for both genders, mirroring the effects observed in the entire sample. Although the differences in the impacts on girls and boys are not large, I find that the positive impacts on math are more pronounced for girls, whereas the impacts on the other outcome variables are stronger for boys. However, given the limited number of observations, the heterogeneity in the treatment effects on these long-term academic outcomes is only suggestive currently.

Robustness checks of the sub-sample: I checked the balance across the three treatment groups and one control group in the sub-sample (see Appendix Table C2). I find that, compared to students in the other groups, the students in the Very Successful role model treatment group had

Appendix Table D1, to access the Senior High School Entrance Exam results of the students recruited for this study in 2023.

Table C1: Treatment impacts on long-term test scores

	All				Girls				Boys			
	(1) Math	(2) Total excl. Math	(3) Total	(4) Below Vocational	(5) Math	(6) Total excl. Math	(7) Total	(8) Below Vocational	(9) Math	(10) Total excl. Math	(11) Total	(12) Below Vocational
T1: Very Successful	0.338**	0.182**	0.189**	-0.097	0.278*	0.217**	0.179*	-0.048	0.441***	0.190**	0.277**	-0.164
	(0.122)	(0.068)	(0.072)	(0.064)	(0.135)	(0.078)	(0.094)	(0.075)	(0.134)	(0.079)	(0.102)	(0.094)
T2: VS-General	0.084	0.094***	0.084**	-0.070	0.040	0.094	0.058	-0.052	0.181	0.103	0.145*	-0.098
	(0.089)	(0.019)	(0.030)	(0.042)	(0.120)	(0.103)	(0.083)	(0.095)	(0.104)	(0.083)	(0.078)	(0.089)
T3: Moderately Successful	0.028	-0.069	-0.069	0.044	-0.038	-0.019	-0.047	0.060	0.156	-0.084	-0.045	-0.008
	(0.113)	(0.052)	(0.066)	(0.044)	(0.128)	(0.041)	(0.047)	(0.068)	(0.113)	(0.129)	(0.138)	(0.045)
Baseline Math	0.536***				0.575***				0.500***			
	(0.116)				(0.101)				(0.145)			
Baseline Total excl. Math		0.504***				0.590***				0.437***		
		(0.098)				(0.069)				(0.134)		
Baseline Total			0.512***	-0.247***			0.596***	-0.251***			0.443***	-0.245***
			(0.100)	(0.055)			(0.071)	(0.052)			(0.136)	(0.066)
Constant	0.188*	0.259***	0.259***	0.344***	0.201	0.215***	0.217***	0.354***	0.122	0.269***	0.266***	0.371***
	(0.103)	(0.024)	(0.024)	(0.044)	(0.127)	(0.034)	(0.034)	(0.069)	(0.101)	(0.066)	(0.063)	(0.048)
Observations	361	361	361	361	186	186	186	186	174	174	174	174
R-squared	0.409	0.349	0.363	0.301	0.421	0.370	0.385	0.300	0.427	0.398	0.408	0.343
Effects of revealing different characteristics of role models:												
BG=*T1-T2*	0.254**	0.088	0.105	-0.027	0.239**	0.123	0.121	0.004	0.260*	0.088	0.131	-0.066
	(0.084)	(0.068)	(0.068)	(0.036)	(0.086)	(0.121)	(0.107)	(0.072)	(0.125)	(0.104)	(0.099)	(0.093)
More successful	0.309***	0.250**	0.259***	-0.141***	0.316***	0.236***	0.226***	-0.108**	0.285**	0.274*	0.321**	-0.156**
=*T1-T3*	(0.089)	(0.091)	(0.078)	(0.036)	(0.083)	(0.072)	(0.070)	(0.036)	(0.113)	(0.149)	(0.130)	(0.068)
control mean	-	-	-	0.435	0.016	-0.064	-0.056	0.426	-0.016	0.067	0.059	0.444
control mean baseline	-0.109	-0.151	-0.148	-	-0.102	-0.073	-0.081	-	-0.116	-0.232	-0.219	-

These long-term Senior High School Entrance Examination test scores of Grade 8 (now graduated) students are provided by three out of the five middle schools. "Total excl. Math" refers to the standardized aggregate score across all subjects excluding Math taken during the Entrance Exam. "Total" refers to the standardized aggregate score across all subjects taken in the Senior High School Entrance Exam. Standardized aggregate scores are composed of subject-standardized scores and normalized. "Below Vocational" is a dummy variable indicating that the student's total test score obtained in the 2024 Senior High School Entrance Exam falls below the cutoff point required for applying to a vocational school. "Baseline Math", "Baseline Total excl. Math" or "Baseline Total" refers to the re-standardized score achieved in the pre-intervention exam. Missing baseline score is replaced by the median pre-intervention exam score, and a dummy variable is included to capture this. Control variables include student baseline test scores and a dummy for the homeroom teacher teaching a science subject for regressions on Math, Total, and Below Vocational School or a dummy for the homeroom teacher teaching the Chinese subject for regressions on Total excl. Math and Total. All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1.

higher pre-intervention average test scores and spent fewer hours studying math during a weekday at the baseline. Students in the Moderately Successful role model treatment group spent more hours studying during a weekday at the baseline than students in the other groups. Therefore, to confirm the robustness of my findings, I include these unbalanced baseline variables as control variables in the regressions (see Appendix Table C3). The results with additional control variables do not reject my findings presented in Appendix Table C1.

Table C2: Balance Test of the Sub-sample — Grade 8 students from the three middle schools

	Control		T1: Very Successful		T2: VS-General		T3: Moderately Successful		Pairwise t-test p-value						F- test p-value	
	N	(1) Mean/(SE)	N	(2) Mean/(SE)	N	(3) Mean/(SE)	N	(4) Mean/(SE)	(1)-(2)	(1)-(3)	(1)-(4)	(2)-(3)	(2)-(4)	(3)-(4)	joint	pooled
Students' Baseline:																
Baseline std Chinese	90	-0.135 (1.317)	106	0.195 (0.937)	61	-0.066 (0.750)	95	-0.026 (0.894)	0.030**	0.691	0.484	0.084*	0.103	0.794	0.111	0.122
Baseline std Math	90	-0.117 (1.156)	107	0.196 (0.820)	61	-0.149 (1.027)	94	0.006 (0.969)	0.028**	0.856	0.420	0.024**	0.155	0.347	0.080	0.177
Baseline std Total	90	-0.172 (1.224)	107	0.230 (0.870)	61	-0.066 (0.819)	95	-0.040 (0.983)	0.006***	0.538	0.395	0.047**	0.048**	0.868	0.032	0.054
Baseline re-std Chinese[a]	90	-0.108 (1.211)	106	0.209 (0.879)	61	-0.070 (0.739)	95	-0.012 (0.860)	0.031**	0.823	0.523	0.058*	0.095*	0.696	0.102	0.149
Baseline re-std Math[a]	90	-0.116 (1.131)	107	0.194 (0.805)	61	-0.151 (1.019)	94	0.005 (0.954)	0.028**	0.841	0.423	0.024**	0.155	0.340	0.079	0.180
Baseline re-std Total[a]	90	-0.155 (1.182)	107	0.240 (0.841)	61	-0.063 (0.809)	95	-0.029 (0.961)	0.006***	0.586	0.407	0.040**	0.046**	0.826	0.031	0.059
Female student	92	0.511 (0.253)	108	0.481 (0.252)	63	0.508 (0.254)	97	0.567 (0.248)	0.680	0.972	0.442	0.740	0.223	0.467	0.673	0.898
Is sibling	92	0.533 (0.252)	105	0.448 (0.250)	63	0.587 (0.246)	97	0.546 (0.250)	0.236	0.504	0.850	0.080*	0.162	0.613	0.301	0.797
Is sibling in G7/G8 same school	92	0.054 (0.052)	108	0.065 (0.061)	64	0.078 (0.073)	97	0.093 (0.085)	0.758	0.554	0.316	0.742	0.459	0.749	0.761	0.449
Boys better math	26	1.000 (0.000)	35	0.971 (0.029)	48	0.917 (0.078)	26	0.885 (0.106)	0.393	0.134	0.077*	0.306	0.181	0.658	0.245	0.157
Hours weekday all	27	2.185 (1.695)	30	1.933 (1.237)	50	1.960 (1.713)	23	2.739 (2.383)	0.434	0.473	0.175	0.926	0.032**	0.029**	0.094	0.838
Hours weekend all	27	2.296 (1.832)	31	2.161 (0.873)	50	2.100 (1.480)	23	2.304 (2.312)	0.657	0.518	0.984	0.811	0.671	0.540	0.881	0.620
Hours weekday math	24	1.708 (1.085)	35	1.114 (0.104)	48	1.646 (1.127)	26	1.808 (1.202)	0.002***	0.813	0.744	0.005***	0.001***	0.538	0.015	0.369
Hours weekend math	24	1.708 (0.998)	35	1.514 (0.375)	48	1.688 (1.156)	26	1.731 (1.165)	0.359	0.937	0.940	0.394	0.325	0.869	0.791	0.760
Aspire above median in Zhong-Kao[b]	26	0.808 (0.162)	35	0.829 (0.146)	48	0.792 (0.168)	25	0.680 (0.227)	0.837	0.872	0.305	0.678	0.186	0.300	0.558	0.742
Homeroom teachers' Baseline:																
Homeroom teacher female	61	1.000 (0.000)	108	1.000 (0.000)	64	0.484 (0.254)	97	1.000 (0.000)	SAME	0.000***	SAME	0.000***	SAME	0.000***	0.000	0.004
Homeroom teacher teaching sci	61	0.000 (0.000)	108	0.657 (0.227)	64	0.516 (0.254)	97	0.381 (0.238)	0.000***	0.000***	0.000***	0.067*	0.000***	0.094*	0.000	0.000

Pairwise t-test displays the p-value from t-tests of the equality of the coefficients between the pairs of columns. Joint F-tests show the p-value from an F-test of equality of the means across all four groups for each covariate. Pooled F-test shows the p-value from a test for pooled assignment to any treatment group versus to the control group.

[a] Re-standardized test scores are standardized test scores based on the observations in Grade 8 (now Grade 9) classes of the middle schools which have provided the long-term Zhong-Kao[b] test scores to me. These will be adjusted later when I receive the long-term Zhong-Kao test scores of Grade 7 students the next year.

[b] Zhong-Kao refers to the Senior High School Entrance Examination in China.

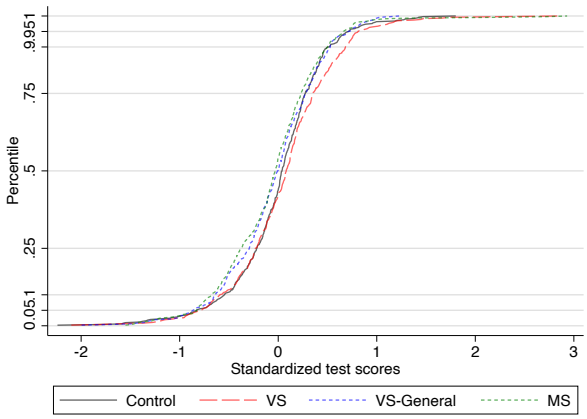
Table C3: Treatment impacts on long-term test scores — with additional control variables

	All				Girls				Boys			
	(1) Math	(2) Total excl. Math	(3) Total	(4) Below Vocational	(5) Math	(6) Total excl. Math	(7) Total	(8) Below Vocational	(9) Math	(10) Total excl. Math	(11) Total	(12) Below Vocational
T1: Very Successful	0.302***	0.162**	0.150*	-0.083	0.250***	0.216**	0.156	-0.049	0.418***	0.185**	0.239**	-0.131
	(0.064)	(0.070)	(0.074)	(0.047)	(0.079)	(0.096)	(0.106)	(0.066)	(0.103)	(0.074)	(0.103)	(0.078)
T2: VS-General	-0.024	0.070*	0.042	-0.013	-0.098	0.104	0.038	0.000	0.136	0.088	0.115	-0.033
	(0.040)	(0.035)	(0.032)	(0.026)	(0.102)	(0.109)	(0.087)	(0.085)	(0.102)	(0.081)	(0.093)	(0.085)
T3: Moderately Successful	0.014	-0.064*	-0.066*	0.046	-0.037	-0.037	-0.056	0.064	0.097	-0.103	-0.077	0.026
	(0.042)	(0.035)	(0.036)	(0.040)	(0.075)	(0.097)	(0.099)	(0.064)	(0.058)	(0.073)	(0.085)	(0.043)
Baseline Math	0.530***				0.583***				0.503***			
	(0.116)				(0.102)				(0.145)			
Baseline Total excl. Math		0.500***				0.591***				0.434***		
		(0.098)				(0.062)				(0.134)		
Baseline Total			0.508***	-0.244***			0.597***	-0.254***			0.444***	-0.240***
			(0.100)	(0.054)			(0.064)	(0.048)			(0.137)	(0.065)
Constant	0.394***	0.300***	0.320***	0.232***	0.429***	0.165*	0.206**	0.321***	0.202	0.334**	0.314*	0.183
	(0.093)	(0.058)	(0.062)	(0.027)	(0.081)	(0.079)	(0.072)	(0.083)	(0.169)	(0.148)	(0.153)	(0.128)
Observations	361	361	361	361	186	186	186	186	174	174	174	174
R-squared	0.425	0.355	0.370	0.315	0.457	0.382	0.400	0.321	0.441	0.410	0.420	0.361
Effects of revealing different characteristics of role models:												
BG=*T1-T2*	0.325***	0.092	0.108	-0.070**	0.348***	0.113	0.117	-0.049	0.282**	0.097	0.124	-0.099
	(0.069)	(0.074)	(0.071)	(0.029)	(0.098)	(0.115)	(0.095)	(0.079)	(0.123)	(0.094)	(0.095)	(0.077)
More successful	0.287***	0.227***	0.216***	-0.130***	0.287***	0.253**	0.211*	-0.113*	0.321***	0.288**	0.316**	-0.157**
=*T1-T3*	(0.035)	(0.071)	(0.063)	(0.021)	(0.059)	(0.105)	(0.095)	(0.052)	(0.081)	(0.113)	(0.114)	(0.052)
control mean	-	-	-	0.435	0.016	-0.064	-0.056	0.426	-0.016	0.067	0.059	0.444
control mean baseline	-0.109	-0.151	-0.148	-	-0.102	-0.073	-0.081	-	-0.116	-0.232	-0.219	-

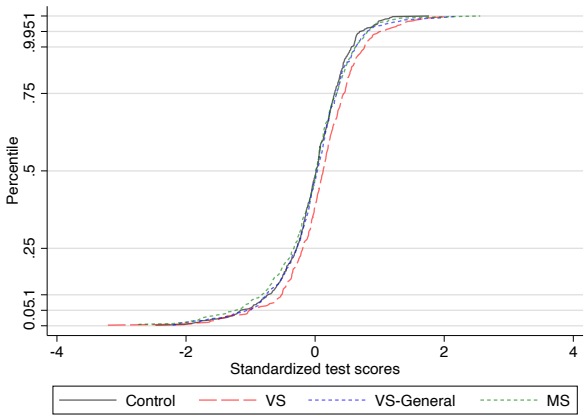
These long-term Senior High School Entrance Examination test scores of Grade 8 (now graduated) students are provided by three out of the five middle schools. "Total excl. Math" refers to the standardized aggregate score across all subjects excluding Math taken during the Entrance Exam. "Total" refers to the standardized aggregate score across all subjects taken in the Senior High School Entrance Exam. Standardized aggregate scores are composed of subject-standardized scores and normalized. "Below Vocational" is a dummy variable indicating that the student's total test score obtained in the 2024 Senior High School Entrance Exam falls below the cutoff point required for applying to a vocational school. "Baseline Math", "Baseline Total excl. Math" or "Baseline Total" refers to the re-standardized score achieved in the pre-intervention exam. Any missing baseline score is replaced by the median pre-intervention exam score, and a dummy variable is included to capture this. Control variables include student baseline test score, overall study hours during a weekday, hours spent studying math during a weekday, and a dummy for the homeroom teacher teaching a science subject for regressions on Math, Total, and Below Vocational School or a dummy for the homeroom teacher teaching the Chinese subject for regressions on Total excl. Math and Total. All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1.

D Additional Figures and Tables

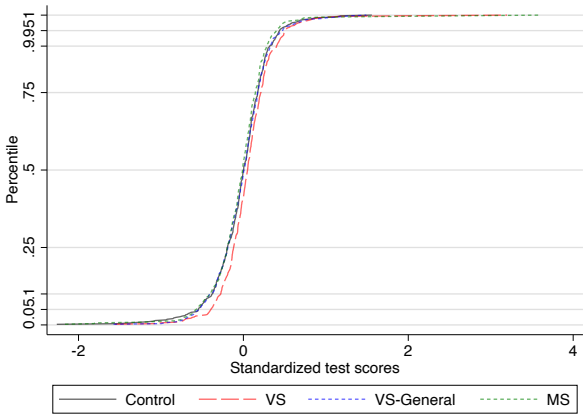
Figure D1: CDFs of math and total test scores



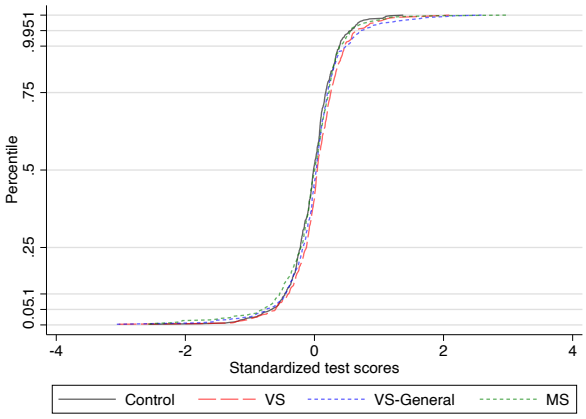
(a) Midterm math test score.



(b) Final math test score.



(c) Midterm total test score.



(d) Final total test score.

VS represents the Very Successful role model treatment, and MS presents the Moderately Successful role model treatment. These figures are plotted using the residuals from regressions where the post-intervention test scores are regressed on baseline test scores, a dummy variable for the student performing below the median within grade 7/8 at her school in the baseline exam, and a dummy variable for the student studying in Grade 8. All regressions include school fixed effects. Errors are clustered at the level of randomization (class level).

Table D1: Information about the Five Recruited Schools

Middle Schools	Number of classes				Students in Grade 7&8		
	Grade 7	Grade 8	Grade 9	Total	#Total	#Participated	#Classes
School 1	10	9	10	29	884	851	19
School 2	4	4	3	11	318	313	8
School 3	3	3	2	8	216	205	6
School 4	6	4	4	14	320	311	10
School 5	3	3	3	9	252	240	6
Total	26	23	22	71	1990	1920	49

These recruited middle schools are close to each other geographically. Enrolled students are randomly assigned to classes. Each class within the middle schools contains 40-50 students.

Table D2: Variable definitions used in the student surveys

Variable Name	Details	Which Survey?
Hours Weekday Math	*How many hours the student spent on studying math on a weekday on average (school lectures excluded)*: $= 1$ if ≤ 2 hours/day; $= 2$ if $2 - 3$ hours/day; $= 3$ if $3 - 4$ hours/day; $= 4$ if $4 - 5$ hours/day; or, $= 5$ if > 5 hours/day.	Student baseline & Student follow-up
Hours Weekend Math	*How many hours the student spent on studying math on a Saturday or Sunday on average*: $= 1$ if ≤ 3 hours/day; $= 2$ if $3 - 5$ hours/day; $= 3$ if $5 - 7$ hours/day; $= 4$ if $7 - 9$ hours/day; or, $= 5$ if > 9 hours/day.	Student baseline & Student follow-up

To be continued. Please refer to the next page.

Table D2: Variable definitions used in the student surveys, continued.

Variable Name	Details	Which Survey?
Hours Weekday All	*How many hours the student spent on studying overall on a weekday on average (school lectures excluded)*: $= 1$ if ≤ 2 hours/day; $= 2$ if $2 - 3$ hours/day; $= 3$ if $3 - 4$ hours/day; $= 4$ if $4 - 5$ hours/day; or, $= 5$ if > 5 hours/day.	Student baseline & Student follow-up
Hours Weekend All	*How many hours the student spent on studying overall on a Saturday or Sunday on average*: $= 1$ if ≤ 3 hours/day; $= 2$ if $3 - 5$ hours/day; $= 3$ if $5 - 7$ hours/day; $= 4$ if $7 - 9$ hours/day; or, $= 5$ if > 9 hours/day.	Student baseline & Student follow-up
Academic aspiration	*The student's aspired ranking in her class for the Senior High School Entrance Examination:* (for classes with around 50 students in total) $= 1$ if Top 5; $= 2$ if Top $6 - 10$; $= 3$ if Top $11 - 15$; $= 4$ if 16−median; $= 5$ if around median; $= 6$ if below the median; or, $= 7$ if "I don't know; Or, I don't have specific goals.".	Student baseline & Student follow-up
Depression Dummy	*Whether the student was experiencing or had experienced heavy depression in the past two weeks*: $= 1$ if *Yes*; $= 0$ if *No*.	Student follow-up
Stress Dummy	*Whether the student was experiencing or had experienced heavy study pressure in the past two weeks*: $= 1$ if *Yes*; $= 0$ if *No*.	Student follow-up
Poor Mental Health Index	A weighted average of the Depression Dummy and the Stress Dummy, following the method of Anderson (2008).	-
Belief	*Did the student agree or disagree with the statement that "Boys are inherently better at math than girls"?* Choose from: *Totally disagree, Slightly disagree, Neutral, Slightly agree*, or *Totally agree*.	Student baseline
Belief	*Which group in the following did the student think has the advantages of studying math at the middle school level?* Choose from: *Boys, Indifferent*, or *Girls*.	Student follow-up

Table D3: Math and Total Test Scores by Gender

	(1) Baseline Math	(2) Baseline Total	(3) CEPS Math	(4) CEPS Total
Student Female	-0.088	0.020	0.132***	0.504***
	(0.055)	(0.052)	(0.017)	(0.016)
Constant	0.053	0.012	-0.068	-0.259***
	(0.159)	(0.165)	(0.062)	(0.078)
Observations	1,853	1,853	18,944	18,944
R-squared	0.002	0.001	0.004	0.063

"Baseline Math" or "Baseline Total" refers to the baseline test scores collected by this field experiment. "CEPS Math" or "CEPS Total" refers to the test scores collected by the China Education Panel Survey 2013-14 (Round 1). Math test scores are standardized. Standardized aggregate scores are composed of subject-standardized scores and normalized. All regressions include the school fixed effects and grade fixed effects. Standard errors in parentheses are clustered at the class level. *** p<0.01, ** p<0.05, * p<0.1

Table D4: Balance Test

	Control		T1: Very Successful		T2: VS-General		T3: Moderately Successful		Pairwise t-test p-value						F-test p-value	
	(1)		(2)		(3)		(4)		(1)-(2)	(1)-(3)	(1)-(4)	(2)-(3)	(2)-(4)	(3)-(4)		
	N	Mean/(SE)	N	Mean/(SE)	N	Mean/(SE)	N	Mean/(SE)							joint	pooled
Students' Baseline:																
Baseline std Chinese	497	-0.034 (0.047)	460	0.024 (0.045)	462	0.021 (0.046)	451	-0.009 (0.046)	0.378	0.411	0.707	0.962	0.613	0.653	0.786	0.381
Baseline std Math	493	0.019 (0.047)	460	0.056 (0.043)	460	-0.032 (0.046)	450	-0.047 (0.048)	0.564	0.437	0.325	0.166	0.112	0.826	0.376	0.613
Baseline std Total	498	-0.011 (0.047)	462	0.008 (0.045)	464	0.016 (0.043)	451	-0.012 (0.048)	0.773	0.674	0.993	0.895	0.767	0.667	0.965	0.769
Grade 8	509	0.438 (0.247)	471	0.484 (0.250)	477	0.407 (0.242)	463	0.497 (0.251)	0.150	0.319	0.067	0.017	0.699	0.006	0.019	0.352
Student female	509	0.483 (0.022)	471	0.486 (0.023)	475	0.488 (0.023)	463	0.497 (0.023)	0.928	0.873	0.675	0.946	0.747	0.799	0.979	0.783
Is sibling	501	0.407 (0.022)	463	0.380 (0.023)	469	0.414 (0.023)	458	0.397 (0.023)	0.391	0.838	0.757	0.296	0.592	0.615	0.742	0.694
Is sibling in G7/G8 same school	509	0.039 (0.009)	471	0.036 (0.009)	477	0.044 (0.009)	463	0.050 (0.010)	0.793	0.710	0.432	0.534	0.306	0.682	0.748	0.705
Boys better math	215	0.586 (0.244)	172	0.529 (0.251)	162	0.580 (0.245)	186	0.645 (0.230)	0.263	0.910	0.226	0.349	0.026	0.216	0.172	0.990
Hours weekday all	217	2.392 (0.089)	158	2.316 (0.095)	150	2.020 (0.111)	185	2.405 (0.104)	0.571	0.009	0.920	0.043	0.534	0.012	0.031	0.223
Hours weekend all	215	2.256 (0.086)	155	2.200 (0.088)	152	2.000 (0.093)	183	2.230 (0.092)	0.657	0.048	0.835	0.119	0.818	0.083	0.200	0.276
Hours weekday math	210	1.848 (0.074)	165	1.648 (0.076)	157	1.726 (0.091)	184	1.679 (0.081)	0.066	0.298	0.126	0.512	0.783	0.700	0.277	0.064
Hours weekend math	211	1.768 (0.065)	167	1.707 (0.066)	154	1.643 (0.077)	185	1.665 (0.068)	0.515	0.216	0.277	0.529	0.662	0.831	0.573	0.206
Aspire above median in Zhong-Kao[a]	216	0.819 (0.026)	168	0.798 (0.031)	164	0.860 (0.027)	186	0.876 (0.024)	0.590	0.293	0.116	0.134	0.044	0.648	0.162	0.382
Homeroom teachers' Baseline:																
Homeroom teacher female	369	0.867 (0.018)	215	1.000 (0.000)	430	0.921 (0.013)	315	0.876 (0.019)	0.000	0.013	0.727	0.000	0.000	0.043	0.000	0.001
Homeroom teacher science subjects	369	0.729 (0.023)	215	0.828 (0.026)	430	0.460 (0.024)	315	0.375 (0.027)	0.006	0.000	0.000	0.000	0.000	0.019	0.000	0.000

Joint F-test p-value by treatment assignment[b]

	T1: Very Successful	T2: VS-General	T3: Moderately Successful	Any Treatment
Students' & Homeroom Teachers' baseline characteristics	0.074	0.083	0.495	0.484
Only Students' baseline characteristics	0.339	0.387	0.876	0.673

"Pairwise t-test" displays the p-value from t-tests of the equality of the coefficients between the pairs of columns. The "Joint F-test" on the right shows the p-value from an F-test of equality of the means across all four groups for each covariate. "Pooled F-test" shows the p-value from a pooled assignment test for any treatment group versus the control group.
[a] Zhong-Kao refers to the Senior High School Entrance Examination in China.
[b] These joint F-test p-values come from regressing the treatment variable on the baseline characteristics and testing if they are jointly zero.

Table D5: Attrition Balance Test

	(1) Attrition Survey-Based	(2) Attrition Exam-Based
T1: Very Successful	0.003	0.028
	(0.033)	(0.041)
T2: VS-General	-0.007	-0.006
	(0.023)	(0.028)
T3: Moderately Successful	-0.022	-0.012
	(0.024)	(0.029)
Observations	1,920	1,990
R-squared	0.009	0.013
T1-T2	0.009	0.035
	(0.028)	(0.035)
T1-T3	0.025	0.040
	(0.028)	(0.036)
T2-T3	0.016	0.005
	(0.017)	(0.022)
Mean	0.085	0.117
Control mean	0.094	0.119
jointly test p-value	0.676	0.726

"Attrition" is a dummy variable for missing the follow-up student survey. Column (1) includes all students that have at least one baseline survey completed. Column (2) includes all students who have taken at least one exam during the Spring 2023 semester based on test score data provided by the middle schools. All regressions included the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1.

Table D6: Correlates of attrition with baseline characteristics

	(1) Attrition
T1: Very Successful	0.017
	(0.015)
T2: VS-General	-0.004
	(0.015)
T3: Moderately Successful	-0.026*
	(0.014)
Baseline std Chinese	-0.021***
	(0.007)
Baseline std Math	-0.012*
	(0.006)
Baseline std Total	-0.018**
	(0.007)
Grade 8	0.018
	(0.013)
Student female	0.009
	(0.013)
Is sibling	-0.010
	(0.012)
Is sibling in G7/G8 same school	-0.089***
	(0.007)
Boys better math	-0.021
	(0.022)
Hours weekday all	0.005
	(0.009)
Hours weekend all	0.007
	(0.009)
Hours weekday math	-0.002
	(0.010)
Hours weekend math	0.004
	(0.010)
Aspire above the median in Zhong-Kao	-0.006
	(0.029)
Homeroom teacher female	0.027
	(0.024)
Homeroom teacher science subjects	0.031**
	(0.016)
Observations	1920
F-test p-value	
With all variables above	0.048
With only students' baseline characteristics	0.169

Linear regression of baseline characteristics on a variable equal to one if the student did not submit their follow-up survey. Each row represents a separate regression. The F-test p-value comes from regressing the attrition variable on the described variables and testing if they are jointly zero. Robust standard errors in parentheses.

Table D7: Scope for Spillover Effects – Evidence from the Student Survey

	All	Girls	Boys	Diff. Girls-Boys
	(1)	(2)	(3)	(4)
Panel A. Siblings Studying at Your School?				
Control Group				
Yes, in Grade 7 or 8.	0.039	0.024	0.053	-0.028*
	(0.194)	(0.155)	(0.225)	[0.076]
Yes, in the same grade as I am.	0.031	0.012	0.049	-0.037***
	(0.175)	(0.110)	(0.217)	[0.008]
Observations	509	246	263	
Any Treatment Groups				
Yes, in Grade 7 or 8.	0.043	0.038	0.049	-0.009
	(0.204)	(0.190)	(0.215)	[0.475]
Yes, in the same grade as I am.	0.032	0.023	0.040	-0.016
	(0.176)	(0.151)	(0.197)	[0.161]
Observations	1,409	691	718	
Panel B. Having Discussed Your Aspirations and Future Goals?				
Control Group				
with family members	0.609	0.626	0.593	0.028
	(0.488)	(0.485)	(0.492)	[0.594]
with classmates	0.324	0.358	0.293	0.066
	(0.469)	(0.480)	(0.456)	[0.112]
with other students from the school	0.147	0.142	0.152	-0.009
	(0.355)	(0.350)	(0.360)	[0.764]
Observations	509	246	263	
Any Treatment Groups				
with family members	0.636	0.641	0.631	0.008
	(0.481)	(0.480)	(0.483)	[0.758]
with classmates	0.356	0.384	0.330	0.056**
	(0.479)	(0.487)	(0.471)	[0.050]
with other students from the school	0.173	0.187	0.160	0.029
	(0.379)	(0.390)	(0.367)	[0.160]
Observations	1,409	691	718	

The summary statistics are computed from the follow-up student survey post-intervention. "Control Group" includes all the students who received no role model treatments. "Any Treatment Groups" include all the students who belonged to the Very Successful, VS-General, or Moderately Successful role model treatment groups. Columns (1) - (3) report average values for all respondents, for girls and boys, respectively. The standard errors are in parentheses. Column (4) reports the within-class difference between girls and boys, which is obtained from a regression of the variable of interest on a dummy variable for the female student. All regressions include school fixed effects and a dummy variable for the student studying in Grade 8. Standard errors are clustered at the unit of randomization (class level). The p-value is reported in square brackets. *** p<0.01, ** p<0.05, * p<0.1

Table D8: Treatment effects – a permutation test

	(1) Midterm Math	(2) Final Math	(3) Midterm Total	(4) Final Total	(5) Poor Mental Health Index	(6) Depression Dummy	(7) Stress Dummy
T1: Very Successful	0.076	0.184	0.073	0.085	0.004	0.003	0.000
	(0.076)	(0.069)	(0.027)	(0.030)	(0.115)	(0.051)	(0.051)
	[0.324]	[0.010]**	[0.009]***	[0.007]***	[0.972]	[0.950]	[0.996]
	{0.015}**	{0.000}***	{0.070}*	{0.020}**	{0.977}	{0.929}	{0.890}
T2: VS-General	-0.045	0.021	0.014	0.044	-0.013	-0.022	0.010
	(0.071)	(0.060)	(0.026)	(0.034)	(0.098)	(0.046)	(0.042)
	[0.533]	[0.732]	[0.601]	[0.203]	[0.893]	[0.641]	[0.809]
	{0.004}***	{0.136}	{0.411}	{0.740}	{0.857}	{0.757}	{0.532}
T3: Moderately Successful	-0.066	0.001	-0.018	-0.011	-0.252	-0.120	-0.093
	(0.074)	(0.062)	(0.029)	(0.038)	(0.077)	(0.033)	(0.037)
	[0.380]	[0.993]	[0.529]	[0.772]	[0.002]***	[0.001]***	[0.016]**
	{0.079}*	{0.616}	{0.479}	{0.709}	{0.000}***	{0.000}***	{0.002}***
Observations	1,857	1,835	1,857	1,835	1,726	1,726	1,726
R-squared	0.742	0.636	0.862	0.771	0.022	0.024	0.018
Effects of being exposed to role model characteristics: *p-values reported*							
T1=T2 (BG)	[0.175]	[0.029]**	[0.070]*	[0.178]	[0.886]	[0.653]	[0.844]
T1=T3 (More successful)	[0.134]	[0.013]**	[0.008]***	[0.007]***	[0.025]***	[0.013]**	[0.075]*

All regressions include the school fixed effects. All regressions control for baseline variables (if any), a dummy variable indicating that the student is in Grade 8, a dummy variable for being a female homeroom teacher, and a dummy variable for the homeroom teacher teaching a science subject. Columns (5)-(7) also control for a dummy for the homeroom teacher teaching Chinese literature. Standard errors in parentheses are clustered at the unit of randomization (class level). Robust p-values in square brackets.

Permutations tests are performed using the `permute` command in Stata. Permutation p-values are shown in curly brackets and are calculated using 10,000 permutations. *** p<0.01, ** p<0.05, * p<0.1.

Table D9: Treatment Effects on the Test Scores and Mental Health

	(1) Midterm Math	(2) Final Math	(3) Midterm Total	(4) Final Total	(5) Poor Mental Health Index	(6) Depression Dummy	(7) Stress Dummy
Control for the same control variables as in Table 2, without Grade FE (i.e., excluding "Grade8" variable):							
T1: Very Successful	0.080	0.160**	0.062**	0.060*	0.001	-0.009	0.010
	(0.078)	(0.072)	(0.027)	(0.034)	(0.111)	(0.049)	(0.052)
T2: VS-General	-0.048	0.043	0.025	0.067	-0.011	-0.013	0.003
	(0.071)	(0.060)	(0.029)	(0.042)	(0.099)	(0.049)	(0.042)
T3: Moderately Successful	-0.065	-0.005	-0.022	-0.018	-0.252***	-0.120***	-0.093**
	(0.074)	(0.063)	(0.031)	(0.040)	(0.077)	(0.035)	(0.037)
BG=*T1-T2*	0.128	0.118	0.037	-0.007	0.013	0.004	0.007
	(0.091)	(0.079)	(0.033)	(0.041)	(0.115)	(0.057)	(0.047)
More Successful=*T1-T3*	0.144	0.166**	0.083**	0.078*	0.254**	0.111**	0.102**
	(0.093)	(0.080)	(0.035)	(0.040)	(0.108)	(0.049)	(0.051)
R-squared	0.742	0.633	0.861	0.767	0.021	0.021	0.016
Control for baseline value of the outcome only, with Grade FE (i.e., including "Grade8" variable):							
T1: Very Successful	0.055	0.145*	0.074***	0.077**	-0.017	-0.019	0.005
	(0.078)	(0.076)	(0.026)	(0.030)	(0.116)	(0.052)	(0.052)
T2: VS-General	-0.054	0.027	0.010	0.047	-0.011	-0.014	0.005
	(0.069)	(0.058)	(0.027)	(0.035)	(0.091)	(0.045)	(0.039)
T3: Moderately Successful	-0.074	-0.017	-0.012	-0.008	-0.211**	-0.105**	-0.074
	(0.069)	(0.065)	(0.030)	(0.041)	(0.098)	(0.045)	(0.045)
BG=*T1-T2*	0.110	0.118	0.064**	0.030	-0.006	-0.005	0.000
	(0.088)	(0.077)	(0.028)	(0.023)	(0.101)	(0.047)	(0.044)
More Successful=*T1-T3*	0.129	0.162*	0.087***	0.085**	0.195*	0.085*	0.079
	(0.088)	(0.083)	(0.031)	(0.032)	(0.107)	(0.047)	(0.049)
R-squared	0.739	0.633	0.861	0.771	0.019	0.019	0.015
Control for baseline value of the outcome only, without Grade FE (i.e., excluding "Grade8" variable):							
T1: Very Successful	0.056	0.141*	0.071**	0.071*	-0.017	-0.021	0.006
	(0.078)	(0.079)	(0.029)	(0.041)	(0.116)	(0.053)	(0.054)
T2: VS-General	-0.055	0.029	0.012	0.050	-0.010	-0.012	0.003
	(0.069)	(0.061)	(0.031)	(0.042)	(0.092)	(0.048)	(0.039)
T3: Moderately Successful	-0.073	-0.021	-0.017	-0.015	-0.212**	-0.107**	-0.072
	(0.068)	(0.070)	(0.033)	(0.045)	(0.098)	(0.047)	(0.046)
BG=*T1-T2*	0.111	0.112	0.059*	0.021	-0.007	-0.009	0.003
	(0.088)	(0.081)	(0.033)	(0.040)	(0.101)	(0.050)	(0.044)
More Successful=*T1-T3*	0.129	0.162*	0.088**	0.085*	0.195*	0.086*	0.078
	(0.087)	(0.088)	(0.036)	(0.044)	(0.107)	(0.048)	(0.050)
R-squared	0.739	0.630	0.860	0.766	0.019	0.017	0.013
Observations	1,857	1,835	1,857	1,835	1,726	1,726	1,726

"Midterm Total" or "Final Total" refers to the standardized aggregate score across all subjects taken in the midterm or final exam. Standardized aggregate scores are composed of subject-standardized scores and normalized. "Poor Mental Heath Index" is a standardized weighted average of "Depression Dummy" and "Stress Dummy", following Anderson (2008). All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1

Table D10: Treatment effects on aggregate test scores excluding Math

	Midterm Aggregate test score excl. Math			Final Aggregate test score excl. Math		
	(1) All students	(2) Girls	(3) Boys	(4) All students	(5) Girls	(6) Boys
T1: VS	0.078***	0.087***	0.071*	0.073**	0.076**	0.070*
	(0.027)	(0.030)	(0.036)	(0.028)	(0.033)	(0.038)
	[0.035]**	[0.031]**	[0.259]	[0.069]*	[0.138]	[0.377]
T2: VS-General	0.008	0.031	-0.012	0.036	0.034	0.037
	(0.023)	(0.030)	(0.037)	(0.032)	(0.045)	(0.042)
	[0.999]	[0.864]	[0.999]	[0.707]	[0.997]	[0.691]
T3: MS	-0.016	0.021	-0.053*	-0.016	0.034	-0.070
	(0.027)	(0.037)	(0.031)	(0.040)	(0.056)	(0.051)
	[0.999]	[0.999]	[0.259]	[0.999]	[0.997]	[0.485]
Observations	1,857	907	949	1,835	904	930
R-squared	0.846	0.849	0.847	0.758	0.748	0.774
control mean		-0.057	0.068		-0.072	0.083
Effects of revealing different role model characteristics:						
BG=*T1-T2*	0.070**	0.056	0.083*	0.036	0.042	0.033
	(0.033)	(0.034)	(0.043)	(0.028)	(0.038)	(0.039)
More Successful=*T1-T3*	0.094***	0.066	0.123***	0.088**	0.042	0.140***
	(0.034)	(0.040)	(0.037)	(0.038)	(0.052)	(0.047)

"Midterm Total" or "Final Total" refers to the standardized aggregate score across all subjects taken in the midterm or final exam. Standardized aggregate scores are composed of subject-standardized scores and normalized. Any missing baseline score is replaced by the median pre-intervention exam score, and a dummy variable is included to capture this. Control variables include student baseline test score, a dummy variable for the student performing below the median within grade 7/8 at her school in the baseline exam, homeroom teacher gender, a dummy variable for the homeroom teacher teaching a science subject, and a dummy variable indicating the student is in Grade 8. All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). Sharpened q-value in square brackets, following the method of Benjamini et al. (2006). *** p<0.01, ** p<0.05, * p<0.1.

Table D11: Treatment effects on the *Depression* and *Stress* variables – logit regressions

	All Students		Girls		Boys	
	(1) Depression	(2) Stress	(3) Depression	(4) Stress	(5) Depression	(6) Stress
T1: Very Successful	0.016	0.002	0.035	0.101	-0.049	-0.100
	(0.215)	(0.227)	(0.268)	(0.266)	(0.253)	(0.248)
T2: VS-General	-0.091	0.040	-0.272	0.337	0.043	-0.296
	(0.194)	(0.184)	(0.267)	(0.266)	(0.187)	(0.184)
T3: Moderately Successful	-0.529***	-0.441**	-0.717***	-0.310	-0.367*	-0.597***
	(0.147)	(0.174)	(0.208)	(0.224)	(0.196)	(0.226)
Constant	0.562*	-0.380	1.398***	-0.181	-0.414	-0.770**
	(0.321)	(0.339)	(0.396)	(0.447)	(0.412)	(0.337)
Observations	1,726	1,726	836	836	890	890
control mean	0.416	0.349	0.511	0.353	0.329	0.346
Effects of revealing different role model characteristics:						
BG=*T1-T2*	0.107	-0.038	0.307	-0.236	-0.092	0.196
	(0.231)	(0.223)	(0.253)	(0.250)	(0.283)	(0.241)
More Successful=*T1-T3*	0.546***	0.443*	0.752***	0.411*	0.318	0.497*
	(0.210)	(0.233)	(0.209)	(0.239)	(0.288)	(0.280)

The Depression (Stress) dummy variable is for the student experiencing or having ever experienced depression (study stress) in the past two weeks. Control variables include baseline total test scores, a dummy indicating whether the student baseline total is below the median or not, homeroom teacher gender, a dummy for the homeroom teacher teaching a science subject, and a dummy for the homeroom teacher teaching Chinese literature or not. All regressions include the school fixed effects and grade fixed effects. Standard errors are clustered at the unit of randomization (class level). Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table D12: Heterogeneous treatment effects: quartiles most and least affected by the **Very Successful** role model treatment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	Midterm Math			Final Math			Midterm Total			Final Total			Poor Mental Health Index		
	least quartile mean	most quartile mean	diff	least quartile mean	most quartile mean	diff	least quartile mean	most quartile mean	diff	least quartile mean	most quartile mean	diff	least quartile mean	most quartile mean	diff
Estimated Effect	0.04	0.19	0.16***	0.07	0.27	0.19***	0.03	0.15	0.11***	0.01	0.12	0.11***	-0.06	0.18	0.23***
Baseline std Chinese	-0.04	-0.33	-0.29***	0.55	-0.55	-1.10***	0.65	-1.11	-1.76***	0.54	-0.64	-1.18***	0.36	-0.18	-0.54***
Baseline std Math	0.26	-0.58	-0.84***	0.64	-0.52	-1.16***	0.80	-1.20	-2.00***	0.63	-0.75	-1.38***	0.23	-0.28	-0.51***
Baseline std Total	0.13	-0.46	-0.59***	0.65	-0.58	-1.23***	0.80	-1.27	-2.07***	0.63	-0.76	-1.39***	0.32	-0.25	-0.57***
Grade 8	0.59	0.41	-0.18***	0.45	0.48	0.03	0.45	0.43	-0.02	0.56	0.41	-0.15***	0.41	0.44	0.03
(Q1) Bottom-performing Girl	0.09	0.23	0.15***	0.00	0.27	0.27***	0.00	0.37	0.37***	0.00	0.29	0.29***	0.00	0.24	0.24***
(Q2) Lower-middle Girl	0.03	0.15	0.12***	0.10	0.08	-0.02	0.03	0.14	0.11***	0.11	0.07	-0.04*	0.18	0.16	-0.02
(Q3) Upper-middle Girl	0.04	0.05	0.01	0.15	0.03	-0.11***	0.13	0.00	-0.12***	0.17	0.03	-0.14***	0.09	0.11	0.03
(Q4) Top-performing Girl	0.17	0.10	-0.08***	0.19	0.12	-0.07***	0.21	0.00	-0.21***	0.18	0.08	-0.09***	0.20	0.02	-0.18***
(Q1) Bottom-performing Boy	0.17	0.20	0.03	0.00	0.29	0.29***	0.00	0.46	0.46***	0.00	0.35	0.35***	0.10	0.08	-0.02
(Q2) Lower-middle Boy	0.10	0.18	0.08***	0.10	0.10	0.00	0.03	0.02	-0.01	0.11	0.07	-0.03*	0.20	0.03	-0.17***
(Q3) Upper-middle Boy	0.18	0.04	-0.14***	0.23	0.03	-0.20***	0.30	0.00	-0.30***	0.21	0.02	-0.19***	0.01	0.33	0.32***
(Q4) Top-performing Boy	0.23	0.05	-0.18***	0.24	0.08	-0.15***	0.30	0.00	-0.30***	0.23	0.07	-0.15***	0.22	0.03	-0.19***
Student female	0.33	0.53	0.20***	0.44	0.50	0.06*	0.37	0.52	0.15***	0.45	0.48	0.02	0.47	0.53	0.06*
Is sibling	0.36	0.45	0.09***	0.35	0.39	0.04	0.40	0.44	0.04	0.33	0.49	0.16***	0.38	0.37	-0.01
Is sibling in G7/G8 same school	0.04	0.05	0.00	0.03	0.04	0.01	0.04	0.04	0.00	0.03	0.06	0.03**	0.03	0.04	0.01
Boys better math	0.96	0.96	0.00	0.96	0.97	0.00	0.97	0.96	-0.02	0.96	0.97	0.01	0.96	0.96	0.00
Aspire above median in Zhong-Kao	0.94	0.93	-0.02	0.95	0.93	-0.02	0.95	0.91	-0.04**	0.95	0.92	-0.03	0.93	0.94	0.01
Hours weekday all	2.04	2.07	0.03	1.97	2.18	0.22***	2.11	2.07	-0.04	2.02	2.13	0.12**	2.12	2.05	-0.07
Hours weekday Chinese	1.21	1.17	-0.05	1.17	1.21	0.04	1.22	1.19	-0.02	1.19	1.25	0.06	1.22	1.16	-0.07
Hours weekday math	1.24	1.18	-0.06	1.23	1.24	0.01	1.23	1.22	-0.01	1.23	1.27	0.04	1.26	1.16	-0.10**
Hours weekend all	2.07	2.00	-0.07	1.99	2.13	0.14***	2.08	2.01	-0.07	2.05	2.04	-0.00	2.06	1.98	-0.07
Hours weekend Chinese	1.21	1.14	-0.07*	1.18	1.20	0.02	1.19	1.17	-0.03	1.19	1.20	0.02	1.19	1.14	-0.05
Hours weekend math	1.24	1.18	-0.07*	1.22	1.26	0.04	1.23	1.21	-0.01	1.23	1.23	0.00	1.23	1.17	-0.06*
Homeroom teacher female	0.94	0.92	-0.02	0.96	0.87	-0.09***	0.95	0.92	-0.04**	0.93	0.92	-0.02	0.94	0.92	-0.02
Homeroom teacher science subjects	0.38	0.41	0.03	0.29	0.73	0.44***	0.42	0.36	-0.06*	0.45	0.38	-0.07**	0.32	0.50	0.18***
Observations	461	460	921	455	455	910	461	460	921	455	455	910	429	428	857

This table shows the mean values of various baseline characteristics of the students most and least affected by the **VS&BG** role model treatment for each of the primary outcomes. The least and most affected quartiles are estimated using conditional treatment effects on the outcome variables measured using causal forest analysis. The diff. shows the differences in the means of that characteristic for the most and least affected quartile, and indicates whether the difference is statistically significant or not.

Table D13: Heterogeneity in treatment effects by baseline quartile (girls)

	(1) Midterm Math	(2) Final Math	(3) Midterm Total	(4) Final Total	(5) Poor Mental Health Index	(6) Depression Dummy	(7) Stress Dummy
T1: Very Successful	0.178	0.397***	0.200***	0.149	0.385**	0.098	0.225**
	(0.147)	(0.147)	(0.073)	(0.098)	(0.150)	(0.064)	(0.094)
T2: VS-General	0.037	0.154	0.081	0.153	-0.029	-0.102	0.076
	(0.148)	(0.160)	(0.074)	(0.147)	(0.202)	(0.099)	(0.091)
T3: Moderately Successful	0.027	0.074	0.051	0.072	-0.248	-0.200**	-0.011
	(0.128)	(0.112)	(0.092)	(0.156)	(0.166)	(0.077)	(0.085)
T1 × Q2	0.038	-0.293	-0.137	-0.118	-0.722***	-0.241**	-0.366***
	(0.176)	(0.184)	(0.095)	(0.113)	(0.189)	(0.093)	(0.104)
T1 × Q3	-0.049	-0.136	-0.101	-0.011	-0.317	-0.010	-0.255*
	(0.144)	(0.149)	(0.106)	(0.137)	(0.287)	(0.121)	(0.142)
T1 × Q4	-0.115	-0.184	-0.191**	-0.073	-0.298	-0.068	-0.181
	(0.152)	(0.163)	(0.087)	(0.120)	(0.291)	(0.126)	(0.155)
T2 × Q2	0.051	-0.164	-0.107	-0.188	0.056	0.017	0.030
	(0.160)	(0.165)	(0.103)	(0.134)	(0.206)	(0.119)	(0.100)
T2 × Q3	-0.114	-0.101	-0.049	-0.032	0.003	0.100	-0.095
	(0.207)	(0.211)	(0.106)	(0.162)	(0.311)	(0.143)	(0.144)
T2 × Q4	-0.050	-0.206	-0.020	-0.189	0.255	0.126	0.089
	(0.187)	(0.212)	(0.090)	(0.234)	(0.298)	(0.125)	(0.146)
T3 × Q2	-0.020	-0.194	-0.137	-0.174	-0.441**	-0.165	-0.207*
	(0.161)	(0.153)	(0.117)	(0.167)	(0.208)	(0.106)	(0.115)
T3 × Q3	-0.216	-0.205	0.029	0.023	0.204	0.253*	-0.078
	(0.159)	(0.166)	(0.122)	(0.209)	(0.312)	(0.128)	(0.152)
T3 × Q4	-0.142	-0.110	0.008	-0.125	0.207	0.121	0.054
	(0.145)	(0.156)	(0.114)	(0.198)	(0.311)	(0.140)	(0.152)
Q2: Lower-middle	0.136	0.541***	0.241**	0.187	0.276	0.060	0.172*
	(0.139)	(0.141)	(0.119)	(0.124)	(0.170)	(0.097)	(0.089)
Q3: Upper-middle	0.372***	0.658***	0.236	-0.005	-0.015	-0.113	0.098
	(0.137)	(0.154)	(0.177)	(0.180)	(0.259)	(0.116)	(0.130)
Q4: Top-performing	0.474***	0.840***	0.304	0.118	-0.070	-0.047	-0.013
	(0.140)	(0.172)	(0.219)	(0.206)	(0.318)	(0.136)	(0.163)
Constant	-0.034	-0.357**	-0.119	0.167	0.443*	0.769***	0.373***
	(0.148)	(0.144)	(0.115)	(0.125)	(0.230)	(0.104)	(0.113)
Observations	831	830	831	830	836	836	836
R-squared	0.758	0.681	0.865	0.757	0.059	0.081	0.040
Overall treatment effects by groups:							
T1 on Q2	0.216*	0.104	0.063	0.032	-0.337**	-0.142	-0.142
	(0.124)	(0.105)	(0.060)	(0.059)	(0.164)	(0.086)	(0.084)
T2 on Q2	0.088	-0.010	-0.026	-0.035	0.027	-0.086	0.106
	(0.108)	(0.116)	(0.077)	(0.072)	(0.232)	(0.118)	(0.105)
T3 on Q2	0.006	-0.120	-0.087	-0.101	-0.689***	-0.365***	-0.218**
	(0.130)	(0.120)	(0.063)	(0.066)	(0.151)	(0.075)	(0.084)
T1 on Q3	0.129	0.261*	0.100*	0.138*	0.068	0.089	-0.030
	(0.104)	(0.134)	(0.059)	(0.072)	(0.277)	(0.118)	(0.124)
T2 on Q3	-0.078	0.054	0.032	0.122*	-0.026	-0.003	-0.019
	(0.123)	(0.125)	(0.056)	(0.070)	(0.197)	(0.099)	(0.085)
T3 on Q3	-0.190	-0.131	0.080	0.096	-0.043	0.054	-0.089
	(0.118)	(0.153)	(0.068)	(0.090)	(0.219)	(0.095)	(0.098)
T1 on Q4	0.063	0.213*	0.009	0.076	0.087	0.030	0.043
	(0.102)	(0.110)	(0.046)	(0.061)	(0.261)	(0.119)	(0.113)
T2 on Q4	-0.013	-0.052	0.061	-0.035	0.226	0.024	0.165
	(0.098)	(0.099)	(0.038)	(0.127)	(0.276)	(0.113)	(0.130)
T3 on Q4	-0.116	-0.036	0.059	-0.053	-0.041	-0.079	0.043
	(0.088)	(0.114)	(0.048)	(0.097)	(0.245)	(0.120)	(0.112)
Effects of revealing different role model characteristics, by groups:							
BG on Q1	0.141	0.242	0.119	-0.004	0.414*	0.201*	0.148
	(0.154)	(0.164)	(0.079)	(0.137)	(0.234)	(0.109)	(0.107)
More Successful on Q1	0.152	0.323**	0.149	0.077	0.633***	0.298***	0.236**
	(0.150)	(0.125)	(0.104)	(0.158)	(0.197)	(0.085)	(0.103)
BG on Q2	0.128	0.114	0.089	0.066	-0.364	-0.056	-0.248***
	(0.137)	(0.132)	(0.066)	(0.069)	(0.218)	(0.111)	(0.088)
More Successful on Q2	0.210	0.224*	0.149***	0.133*	0.352**	0.222***	0.076
	(0.157)	(0.134)	(0.056)	(0.075)	(0.135)	(0.066)	(0.070)
BG on Q3	0.206*	0.207**	0.067	0.017	0.094	0.091	-0.011
	(0.107)	(0.100)	(0.043)	(0.054)	(0.225)	(0.106)	(0.099)
More Successful on Q3	0.319***	0.391***	0.020	0.043	0.111	0.035	0.058
	(0.100)	(0.126)	(0.058)	(0.079)	(0.248)	(0.104)	(0.110)
BG on Q4	0.076	0.265**	-0.052	0.111	-0.139	0.006	-0.122
	(0.119)	(0.115)	(0.041)	(0.135)	(0.213)	(0.098)	(0.096)
More Successful on Q4	0.178*	0.249*	-0.050	0.129	0.128	0.109	0.000
	(0.103)	(0.127)	(0.051)	(0.100)	(0.189)	(0.109)	(0.082)

Quartile refers to the quartile of the pre-intervention exam scores, from Q1 (bottom-performing) to Q4 (top-performing). "Midterm Total" or "Final Total" refers to the standardized aggregate score across all subjects taken in the midterm or final exam. Standardized aggregate scores are composed of subject-standardized scores and normalized. "Mental Heath Index" is a standardized weighted average of "Depression Dummy" and "Stress Dummy", following Anderson (2008). "Baseline: Math" or "Baseline Total" refers to the standardized score achieved in the pre-intervention exam. Missing baseline score is replaced by the median pre-intervention exam score and a dummy variable is included to capture this. Control variables include student gender, student baseline test score, whether the student baseline total is below the median, homeroom teacher gender, a dummy variable for the homeroom teacher teaching a science subject, and a dummy variable indicating the student is in Grade 8. Columns (3)-(7) also control for a dummy for homeroom teachers teaching Chinese literature. All regressions included the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1.

Table D14: Heterogeneity in treatment effects on beliefs, efforts, and aspirations by baseline quartile (girls)

	Belief	Effort Math			Effort All			Aspiration
	boys better math	Weekday Math	Weekend Math	Effort Math	Weekday All	Weekend All	Effort All	Zhong-Kao above median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
T1: Very Successful	-0.066	0.104	0.004	0.543	0.490***	0.169	3.290**	0.054*
	(0.042)	(0.163)	(0.133)	(1.604)	(0.171)	(0.170)	(1.547)	(0.028)
T2: VS-General	-0.078**	0.005	-0.007	0.062	-0.194	-0.169	-1.978	0.002
	(0.038)	(0.138)	(0.131)	(1.386)	(0.232)	(0.186)	(2.043)	(0.035)
T3: Moderately Successful	-0.058	-0.097	0.047	-0.415	-0.222	-0.095	-2.006	-0.000
	(0.043)	(0.157)	(0.138)	(1.643)	(0.172)	(0.132)	(1.444)	(0.037)
T1 × Q2	-0.026	0.001	0.050	0.433	-0.515**	-0.071	-2.980*	-0.050*
	(0.048)	(0.254)	(0.182)	(2.387)	(0.215)	(0.202)	(1.595)	(0.026)
T1 × Q3	0.073	-0.663**	-0.237	-5.078*	-0.955**	-0.416	-7.090**	-0.104**
	(0.070)	(0.309)	(0.238)	(2.925)	(0.361)	(0.301)	(3.158)	(0.040)
T1 × Q4	-0.029	-0.301	-0.223	-2.695	-0.494	-0.239	-3.733	-0.074**
	(0.054)	(0.283)	(0.234)	(2.743)	(0.313)	(0.316)	(3.050)	(0.032)
T2 × Q2	0.008	0.046	0.053	0.679	0.164	0.065	1.402	-0.008
	(0.068)	(0.197)	(0.159)	(1.794)	(0.303)	(0.228)	(2.274)	(0.038)
T2 × Q3	0.129*	-0.456*	-0.175	-3.557	-0.113	-0.054	-0.570	-0.027
	(0.074)	(0.262)	(0.232)	(2.542)	(0.382)	(0.373)	(3.585)	(0.042)
T2 × Q4	0.097	-0.000	0.029	0.210	0.210	0.078	1.868	-0.035
	(0.064)	(0.226)	(0.195)	(2.157)	(0.352)	(0.311)	(3.237)	(0.044)
T3 × Q2	0.023	0.088	-0.095	0.188	0.110	-0.006	1.031	-0.023
	(0.053)	(0.202)	(0.156)	(1.979)	(0.231)	(0.235)	(1.882)	(0.047)
T3 × Q3	0.135*	-0.063	-0.088	-0.571	-0.087	-0.067	-0.264	-0.021
	(0.071)	(0.289)	(0.214)	(2.662)	(0.364)	(0.277)	(3.060)	(0.041)
T3 × Q4	0.099*	-0.092	-0.234	-1.452	0.238	0.051	1.714	-0.016
	(0.059)	(0.206)	(0.179)	(2.076)	(0.337)	(0.260)	(2.920)	(0.041)
Q2: Lower-middle	-0.020	0.044	0.086	0.691	0.069	0.116	0.760	0.017
	(0.032)	(0.141)	(0.088)	(1.202)	(0.194)	(0.196)	(1.540)	(0.024)
Q3: Upper-middle	-0.119**	0.217	0.115	1.734	0.047	0.204	0.880	0.001
	(0.058)	(0.257)	(0.204)	(2.382)	(0.375)	(0.291)	(3.249)	(0.028)
Q4: Top-performing	-0.080	-0.083	0.126	-0.101	-0.211	0.227	-0.278	-0.020
	(0.060)	(0.185)	(0.169)	(1.710)	(0.361)	(0.302)	(3.165)	(0.034)
Constant	0.871***	1.245***	1.139***	10.506***	2.809***	2.186***	25.474***	1.038***
	(0.079)	(0.248)	(0.206)	(2.345)	(0.290)	(0.251)	(2.490)	(0.035)
Observations	836	836	836	836	836	836	836	820
R-squared	0.056	0.100	0.071	0.099	0.080	0.099	0.095	0.053
Overall treatment effects by groups:								
T1 on Q2	-0.092**	0.104	0.055	0.976	-0.024	0.098	0.310	0.005
	(0.039)	(0.202)	(0.134)	(1.840)	(0.171)	(0.215)	(1.680)	(0.010)
T2 on Q2	-0.070	0.051	0.046	0.741	-0.030	-0.104	-0.576	-0.006
	(0.042)	(0.181)	(0.149)	(1.730)	(0.200)	(0.187)	(1.674)	(0.008)
T3 on Q2	-0.035	-0.009	-0.048	-0.227	-0.112	-0.101	-0.974	-0.023
	(0.043)	(0.166)	(0.141)	(1.613)	(0.168)	(0.206)	(1.577)	(0.020)
T1 on Q3	0.006	-0.560**	-0.233	-4.535*	-0.465	-0.247	-3.800	-0.049*
	(0.077)	(0.243)	(0.200)	(2.328)	(0.284)	(0.289)	(2.712)	(0.029)
T2 on Q3	0.051	-0.452**	-0.182	-3.495*	-0.308	-0.223	-2.549	-0.025
	(0.056)	(0.191)	(0.168)	(1.842)	(0.268)	(0.301)	(2.665)	(0.015)
T3 on Q3	0.077	-0.160	-0.042	-0.986	-0.309	-0.162	-2.270	-0.021**
	(0.049)	(0.218)	(0.176)	(2.034)	(0.261)	(0.257)	(2.440)	(0.009)
T1 on Q4	-0.095	-0.198	-0.218	-2.152	-0.004	-0.070	-0.443	-0.019
	(0.071)	(0.170)	(0.145)	(1.673)	(0.251)	(0.243)	(2.411)	(0.014)
T2 on Q4	0.020	0.004	0.023	0.272	0.015	-0.091	-0.111	-0.033
	(0.052)	(0.207)	(0.143)	(1.901)	(0.273)	(0.226)	(2.410)	(0.020)
T3 on Q4	0.041	-0.189	-0.187	-1.867	0.016	-0.044	-0.292	-0.016
	(0.045)	(0.154)	(0.115)	(1.492)	(0.262)	(0.197)	(2.272)	(0.011)
Effects of revealing different role model characteristics, by groups:								
BG on Q1	0.012	0.099	0.011	0.481	0.685***	0.338*	5.269**	0.052**
	(0.056)	(0.184)	(0.164)	(1.861)	(0.249)	(0.197)	(2.136)	(0.025)
More Successful on Q1	-0.008	0.200	-0.042	0.958	0.712***	0.264*	5.296***	0.054**
	(0.059)	(0.177)	(0.152)	(1.868)	(0.186)	(0.150)	(1.569)	(0.026)
BG on Q2	-0.022	0.053	0.009	0.235	0.006	0.203	0.886	0.011
	(0.044)	(0.190)	(0.160)	(1.844)	(0.179)	(0.156)	(1.589)	(0.010)
More Successful on Q2	-0.057	0.113	0.103	1.203	0.088	0.200	1.284	0.028
	(0.045)	(0.173)	(0.146)	(1.704)	(0.152)	(0.190)	(1.598)	(0.023)
BG on Q3	-0.045	-0.108	-0.051	-1.039	-0.157	-0.024	-1.251	-0.025
	(0.067)	(0.166)	(0.181)	(1.806)	(0.222)	(0.267)	(2.260)	(0.030)
More Successful on Q3	-0.070	-0.400**	-0.191	-3.549*	-0.156	-0.085	-1.530	-0.029
	(0.060)	(0.187)	(0.172)	(1.883)	(0.213)	(0.203)	(1.915)	(0.028)
BG on Q4	-0.115*	-0.202	-0.241	-2.424	-0.019	0.021	-0.332	0.014
	(0.067)	(0.203)	(0.167)	(1.974)	(0.225)	(0.272)	(2.330)	(0.022)
More Successful on Q4	-0.136**	-0.009	-0.031	-0.285	-0.020	-0.026	-0.151	-0.003
	(0.056)	(0.152)	(0.137)	(1.564)	(0.194)	(0.240)	(2.035)	(0.013)

Quartile refers to the quartile of the pre-intervention exam scores, from Q1 (bottom-performing) to Q4 (top-performing). "Effort on studying math" or "Effort on studying all subjects" refers to the log of total hours spent during a week on studying math or all subjects. Aspiration is defined as a dummy for "Aspire to rank above median in the Senior High School Entrance Examination". "Belief" is a dummy for believing that boys are inherently better at math than girls. Control variables include student gender, student baseline aggregate test score, student baseline response to each question if exists, homeroom teacher gender, a dummy variable for the homeroom teacher teaching a science subject, and a dummy variable indicating the student is in Grade 8. Regressions on overall efforts and aspirations also control for a dummy for the homeroom teacher teaching Chinese literature. Regressions on efforts spent on studying math and beliefs also control for student pre-intervention math test score. All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1. Visualized marginal role model treatment effects on each quartile of girls are shown in Figure 5.

Table D15: Heterogeneity in treatment impacts by baseline quartile (boys)

	Test Scores				Poor Mental Health			Belief	Effort Math			Effort All			Aspiration
	Midterm Math	Final Math	Midterm Total	Final Total	Index	Depression Dummy	Stress Dummy	boys better math	Weekday Math	Weekend Math	Effort Math	Weekday All	Weekend All	Effort All	Zhong-Kao above median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
T1: Very Successful	0.039	0.315***	0.158*	0.226**	0.037	-0.016	0.047	-0.045	0.248	0.127	2.172	0.198	0.103	1.730	-0.008
	(0.175)	(0.102)	(0.082)	(0.090)	(0.239)	(0.105)	(0.111)	(0.036)	(0.199)	(0.231)	(2.240)	(0.192)	(0.193)	(1.667)	(0.038)
T2: VS-General	-0.137	0.094	0.055	0.153	0.110	0.087	0.006	-0.055	-0.021	0.004	-0.225	-0.002	-0.051	-0.340	0.020
	(0.139)	(0.134)	(0.056)	(0.142)	(0.199)	(0.091)	(0.092)	(0.034)	(0.201)	(0.190)	(2.028)	(0.226)	(0.198)	(2.002)	(0.032)
T3: Moderately Successful	0.006	-0.090	0.003	-0.125	-0.069	-0.035	-0.023	-0.096**	0.032	0.108	0.498	-0.126	-0.057	-0.934	-0.016
	(0.168)	(0.126)	(0.100)	(0.115)	(0.190)	(0.080)	(0.096)	(0.042)	(0.153)	(0.200)	(1.788)	(0.236)	(0.186)	(1.999)	(0.043)
T1 × Q2	0.105	-0.132	-0.065	-0.099	-0.496**	-0.168	-0.249**	0.038	-0.556**	-0.432*	-5.385**	-0.473**	-0.212	-3.795*	-0.013
	(0.146)	(0.146)	(0.097)	(0.115)	(0.211)	(0.118)	(0.098)	(0.055)	(0.252)	(0.241)	(2.597)	(0.232)	(0.248)	(2.025)	(0.053)
T1 × Q3	-0.058	-0.245**	-0.178	-0.197*	0.142	0.076	0.044	0.027	-0.107	-0.010	-0.650	-0.233	0.092	-0.881	-0.007
	(0.193)	(0.118)	(0.107)	(0.113)	(0.305)	(0.128)	(0.152)	(0.042)	(0.314)	(0.330)	(3.403)	(0.238)	(0.301)	(2.383)	(0.046)
T1 × Q4	0.023	-0.242	-0.138	-0.262	-0.064	0.047	-0.099	0.050	-0.443*	-0.237	-3.803	-0.450	-0.376	-4.016*	0.023
	(0.179)	(0.157)	(0.109)	(0.163)	(0.284)	(0.140)	(0.131)	(0.044)	(0.237)	(0.281)	(2.706)	(0.279)	(0.251)	(2.110)	(0.045)
T2 × Q2	0.244**	0.113	0.022	0.040	-0.391	-0.164	-0.165	-0.004	-0.052	-0.335	-1.723	0.065	0.077	0.779	-0.020
	(0.118)	(0.175)	(0.077)	(0.195)	(0.282)	(0.129)	(0.136)	(0.052)	(0.256)	(0.258)	(2.628)	(0.311)	(0.297)	(2.899)	(0.047)
T2 × Q3	0.086	-0.203	-0.045	-0.226	-0.371	-0.201	-0.112	0.036	0.481*	0.227	4.184	0.123	0.321	2.407	-0.017
	(0.172)	(0.185)	(0.082)	(0.171)	(0.337)	(0.154)	(0.156)	(0.061)	(0.259)	(0.273)	(2.715)	(0.291)	(0.317)	(2.846)	(0.033)
T2 × Q4	0.063	-0.140	-0.123	-0.188	-0.049	0.001	-0.042	0.044	0.052	-0.013	0.255	-0.278	-0.042	-1.500	-0.007
	(0.161)	(0.184)	(0.089)	(0.185)	(0.245)	(0.118)	(0.114)	(0.054)	(0.251)	(0.246)	(2.564)	(0.306)	(0.287)	(2.689)	(0.040)
T3 × Q2	-0.054	0.100	0.012	0.136	-0.429*	-0.125	-0.235**	-0.004	-0.176	-0.158	-1.420	0.117	0.048	0.993	0.041
	(0.185)	(0.209)	(0.115)	(0.105)	(0.238)	(0.134)	(0.108)	(0.080)	(0.232)	(0.237)	(2.323)	(0.301)	(0.280)	(2.608)	(0.053)
T3 × Q3	-0.306	0.140	-0.125	0.126	-0.446	-0.216*	-0.161	0.072	0.331	0.090	2.820	0.601*	0.245	4.494	0.011
	(0.227)	(0.166)	(0.155)	(0.138)	(0.288)	(0.123)	(0.140)	(0.053)	(0.249)	(0.270)	(2.621)	(0.305)	(0.301)	(2.757)	(0.047)
T3 × Q4	-0.072	0.192	-0.014	0.127	0.067	0.100	-0.042	0.098	0.023	0.169	1.105	0.005	-0.023	-0.010	0.005
	(0.184)	(0.147)	(0.133)	(0.128)	(0.206)	(0.084)	(0.118)	(0.064)	(0.218)	(0.247)	(2.307)	(0.280)	(0.199)	(2.164)	(0.056)
Q2: Lower-middle	0.177*	0.351***	0.010	0.121	0.567***	0.207**	0.270***	-0.034	0.347**	0.357**	3.613**	0.053	-0.216	-0.807	-0.022
	(0.097)	(0.077)	(0.071)	(0.093)	(0.165)	(0.081)	(0.086)	(0.045)	(0.168)	(0.142)	(1.625)	(0.214)	(0.192)	(1.718)	(0.037)
Q3: Upper-middle	0.470***	0.686***	0.088	0.361***	0.400	0.135	0.201	-0.060	-0.226	-0.160	-2.583	-0.266	-0.468*	-3.991*	-0.013
	(0.119)	(0.110)	(0.096)	(0.116)	(0.295)	(0.121)	(0.143)	(0.046)	(0.232)	(0.266)	(2.553)	(0.217)	(0.265)	(2.025)	(0.030)
Q4: Top-performing	0.539***	0.788***	0.007	0.369**	0.237	-0.004	0.201	-0.082	0.009	-0.172	-1.112	-0.172	-0.223	-2.461	-0.040
	(0.130)	(0.131)	(0.113)	(0.142)	(0.261)	(0.112)	(0.134)	(0.054)	(0.253)	(0.266)	(2.672)	(0.288)	(0.281)	(2.494)	(0.039)
Constant	-0.050	-0.457***	0.060	-0.126	-0.306	0.349***	0.160	0.816***	1.303***	1.268***	11.777***	2.968***	2.795***	28.809***	1.054***
	(0.214)	(0.157)	(0.072)	(0.120)	(0.239)	(0.096)	(0.118)	(0.108)	(0.356)	(0.298)	(3.534)	(0.281)	(0.232)	(2.439)	(0.033)
Observations	881	863	881	863	890	890	890	890	890	890	890	890	890	890	853
R-squared	0.767	0.643	0.883	0.792	0.044	0.035	0.046	0.058	0.066	0.068	0.073	0.094	0.078	0.093	0.042
Overall treatment effects by groups:															
T1 on Q2	0.144	0.183	0.093	0.128	-0.459**	-0.184**	-0.202**	-0.008	-0.307	-0.305	-3.213	-0.274	-0.109	-2.065	-0.021
	(0.129)	(0.155)	(0.072)	(0.088)	(0.179)	(0.090)	(0.083)	(0.040)	(0.267)	(0.272)	(2.800)	(0.294)	(0.183)	(2.400)	(0.034)
T2 on Q2	0.107	0.207**	0.077	0.193**	-0.282	-0.078	-0.159*	-0.059	-0.072	-0.331	-1.947	0.063	0.026	0.440	0.000
	(0.130)	(0.098)	(0.075)	(0.094)	(0.187)	(0.084)	(0.088)	(0.045)	(0.299)	(0.294)	(3.031)	(0.314)	(0.200)	(2.541)	(0.028)
T3 on Q2	-0.047	0.010	0.015	0.011	-0.498**	-0.160	-0.258***	-0.100	-0.144	-0.049	-0.922	-0.009	-0.008	0.059	0.025
	(0.114)	(0.168)	(0.059)	(0.102)	(0.193)	(0.101)	(0.084)	(0.083)	(0.285)	(0.297)	(3.002)	(0.317)	(0.234)	(2.698)	(0.024)
T1 on Q3	-0.019	0.070	-0.020	0.030	0.179	0.060	0.090	-0.019	0.141	0.117	1.522	-0.034	0.196	0.849	-0.015
	(0.101)	(0.078)	(0.049)	(0.066)	(0.190)	(0.082)	(0.095)	(0.039)	(0.236)	(0.217)	(2.452)	(0.199)	(0.161)	(1.597)	(0.022)
T2 on Q3	-0.051	-0.109	0.010	-0.073	-0.261	-0.115	-0.105	-0.019	0.460**	0.230	3.960*	0.122	0.269	2.068	0.003
	(0.092)	(0.101)	(0.061)	(0.066)	(0.211)	(0.102)	(0.100)	(0.050)	(0.220)	(0.195)	(2.176)	(0.218)	(0.214)	(2.085)	(0.009)
T3 on Q3	-0.300**	0.050	-0.122	0.001	-0.515***	-0.251***	-0.184**	-0.024	0.363*	0.198	3.318	0.475***	0.188	3.560**	-0.005
	(0.120)	(0.082)	(0.076)	(0.053)	(0.176)	(0.078)	(0.088)	(0.044)	(0.196)	(0.186)	(1.994)	(0.176)	(0.165)	(1.529)	(0.010)
T1 on Q4	0.062	0.073	0.020	-0.036	-0.027	0.031	-0.053	0.005	-0.195	-0.110	-1.632	-0.252	-0.273	-2.286	0.015
	(0.080)	(0.133)	(0.048)	(0.102)	(0.217)	(0.108)	(0.101)	(0.036)	(0.133)	(0.120)	(1.260)	(0.257)	(0.183)	(1.968)	(0.014)
T2 on Q4	-0.074	-0.046	-0.068	-0.035	0.061	0.088	-0.036	-0.011	0.031	-0.009	0.030	-0.280	-0.094	-1.840	0.013
	(0.102)	(0.102)	(0.056)	(0.080)	(0.168)	(0.084)	(0.074)	(0.044)	(0.166)	(0.149)	(1.594)	(0.208)	(0.224)	(1.908)	(0.016)
T3 on Q4	-0.066	0.102	-0.011	0.002	-0.002	0.065	-0.065	0.002	0.055	0.277**	1.603	-0.121	-0.079	-0.944	-0.012
	(0.092)	(0.089)	(0.054)	(0.052)	(0.155)	(0.070)	(0.076)	(0.044)	(0.156)	(0.133)	(1.473)	(0.211)	(0.153)	(1.675)	(0.023)
Effects of revealing different role model characteristics, by groups:															
BG on Q1	0.176	0.222	0.103	0.074	-0.073	-0.103	0.040	0.010	0.269	0.123	2.396	0.200	0.155	2.070	-0.028
	(0.197)	(0.140)	(0.072)	(0.139)	(0.235)	(0.119)	(0.093)	(0.041)	(0.216)	(0.200)	(2.195)	(0.209)	(0.230)	(2.058)	(0.032)
More Successful on Q1	0.033	0.406***	0.155	0.351***	0.106	0.019	0.070	0.051	0.216	0.019	1.674	0.324	0.160	2.664	0.008
	(0.216)	(0.138)	(0.105)	(0.114)	(0.233)	(0.112)	(0.101)	(0.047)	(0.158)	(0.201)	(1.860)	(0.196)	(0.222)	(1.959)	(0.038)
BG on Q2	0.037	-0.024	0.017	-0.065	-0.177	-0.106	-0.044	0.052	-0.235	0.026	-1.266	-0.338	-0.134	-2.504	-0.021
	(0.154)	(0.150)	(0.072)	(0.098)	(0.167)	(0.074)	(0.085)	(0.047)	(0.216)	(0.194)	(2.143)	(0.219)	(0.162)	(1.828)	(0.033)
More Successful on Q2	0.192	0.173	0.078	0.117	0.039	-0.024	0.056	0.092	-0.164	-0.256	-2.291	-0.265	-0.100	-2.124	-0.046*
	(0.147)	(0.214)	(0.060)	(0.109)	(0.168)	(0.086)	(0.084)	(0.088)	(0.199)	(0.193)	(2.075)	(0.213)	(0.199)	(1.993)	(0.025)
BG on Q3	0.032	0.179	-0.030	0.103	0.440**	0.175*	0.196**	0.001	-0.319	-0.113	-2.438	-0.156	-0.074	-1.219	-0.018
	(0.111)	(0.112)	(0.058)	(0.076)	(0.207)	(0.096)	(0.097)	(0.057)	(0.295)	(0.262)	(3.029)	(0.279)	(0.208)	(2.440)	(0.025)
More Successful on Q3	0.281**	0.021	0.102	0.028	0.694***	0.311***	0.274***	0.006	-0.222	-0.081	-1.796	-0.510**	0.007	-2.711	-0.010
	(0.130)	(0.096)	(0.076)	(0.063)	(0.168)	(0.072)	(0.080)	(0.053)	(0.259)	(0.235)	(2.697)	(0.232)	(0.155)	(1.919)	(0.024)
BG on Q4	0.136	0.120	0.088**	-0.001	-0.088	-0.057	-0.017	0.016	-0.226	-0.101	-1.662	0.028	-0.179	-0.446	0.002
	(0.106)	(0.152)	(0.043)	(0.115)	(0.238)	(0.126)	(0.101)	(0.048)	(0.171)	(0.159)	(1.675)	(0.296)	(0.234)	(2.428)	(0.013)
More Successful on Q4	0.128	-0.028	0.031	-0.038	-0.025	-0.034	0.012	0.003	-0.250*	-0.387***	-3.235**	-0.131	-0.194	-1.341	0.027
	(0.086)	(0.141)	(0.041)	(0.103)	(0.235)	(0.116)	(0.105)	(0.047)	(0.143)	(0.126)	(1.352)	(0.289)	(0.184)	(2.251)	(0.022)

Quartile refers to the quartile of the pre-intervention exam scores, from Q1 (bottom-performing) to Q4 (top-performing). "Midterm Total" or "Final Total" refers to the standardized aggregate score across all subjects taken in the midterm or final exam. Standardized aggregate scores are composed of subject-standardized scores and normalized. "Mental Heath Index" is a standardized weighted average of "Depression Dummy" and "Stress Dummy", following Anderson (2008). "Baseline: Math" or "Baseline Total" refers to the standardized score achieved in the pre-intervention exam. Missing baseline score is replaced by the median pre-intervention exam score and a dummy variable is included to capture this. Control variables include student gender, student baseline test score, whether the student baseline total is below the median, homeroom teacher gender, a dummy variable for the homeroom teacher teaching a science subject, and a dummy variable indicating the student is in Grade 8. Columns (3)-(7) also control for a dummy for homeroom teachers teaching Chinese literature. All regressions included the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1.

Table D16: Heterogeneity in treatment impacts on realized midterm exam rankings

	Girls		Boys	
	(1)	(2)	(3)	(4)
	Below 25% in Class	Exam Ranking in the Grade	Below 25% in Class	Exam Ranking in the Grade
T1: Very Successful	0.051	-41.278	0.151	-6.439
	(0.115)	(29.667)	(0.103)	(32.405)
T2: VS-General	-0.018	-11.316	0.055	8.378
	(0.095)	(28.649)	(0.106)	(29.216)
T3: Moderately Successful	0.029	-26.371	0.085	-5.117
	(0.087)	(32.386)	(0.094)	(32.445)
T1 × Q2	-0.022	30.404	-0.102	-21.609
	(0.125)	(19.433)	(0.140)	(22.091)
T1 × Q3	-0.096	36.004	-0.242	2.236
	(0.142)	(36.782)	(0.150)	(36.811)
T1 × Q4	-0.032	48.824	-0.094	16.429
	(0.119)	(53.117)	(0.109)	(49.436)
T2 × Q2	0.002	34.148*	-0.012	-21.218
	(0.119)	(17.433)	(0.122)	(18.764)
T2 × Q3	-0.009	33.161	-0.177	-5.333
	(0.140)	(31.309)	(0.155)	(31.511)
T2 × Q4	-0.001	29.484	-0.070	-5.826
	(0.102)	(49.562)	(0.104)	(45.879)
T3 × Q2	-0.019	26.962	-0.014	4.160
	(0.106)	(17.994)	(0.118)	(19.372)
T3 × Q3	-0.065	37.519	-0.135	11.925
	(0.134)	(36.293)	(0.150)	(34.906)
T3 × Q4	-0.059	45.613	-0.061	-0.377
	(0.100)	(48.427)	(0.099)	(48.487)
Q2: Lower-middle	-0.232**	-22.892	-0.237**	4.179
	(0.099)	(17.366)	(0.100)	(14.741)
Q3: Upper-middle	-0.038	-62.116*	-0.076	-54.034**
	(0.133)	(31.171)	(0.142)	(26.278)
Q4: Top-performing	0.018	-96.795**	-0.108	-78.068**
	(0.147)	(41.745)	(0.123)	(35.714)
Constant	0.174	258.103***	0.459***	261.012***
	(0.153)	(40.107)	(0.135)	(39.637)
Observations	831	831	881	881
R-squared	0.401	0.794	0.457	0.805
Overall treatment effects by groups:				
T1 on Q2	0.029	-10.874	0.048	-28.048*
	(0.089)	(14.918)	(0.091)	(15.139)
T2 on Q2	-0.016	22.832	0.043	-12.840
	(0.085)	(17.629)	(0.095)	(17.033)
T3 on Q2	0.010	0.591	0.071	-0.956
	(0.080)	(17.752)	(0.109)	(19.393)
T1 on Q3	-0.044	-5.274	-0.091	-4.204
	(0.077)	(12.580)	(0.095)	(12.160)
T2 on Q3	-0.027	21.845	-0.121	3.045
	(0.102)	(13.189)	(0.117)	(14.763)
T3 on Q3	-0.036	11.148	-0.050	6.808
	(0.094)	(10.946)	(0.108)	(11.628)
T1 on Q4	0.020	7.547	0.057**	9.989
	(0.031)	(35.865)	(0.028)	(31.164)
T2 on Q4	-0.020	18.168	-0.015	2.552
	(0.035)	(35.080)	(0.036)	(32.374)
T3 on Q4	-0.030	19.242	0.024	-5.493
	(0.042)	(28.266)	(0.031)	(28.936)
Effects of revealing different characteristics of role models, by groups:				
BG on Q1	0.069	-29.962	0.095	-14.818
	(0.101)	(23.579)	(0.101)	(26.603)
More Successful on Q1	0.022	-14.907	0.066	-1.323
	(0.094)	(28.342)	(0.088)	(31.384)
BG on Q2	0.045	-33.706**	0.005	-15.208
	(0.088)	(14.065)	(0.099)	(11.157)
More Successful on Q2	0.019	-11.465	-0.023	-27.092*
	(0.082)	(15.477)	(0.113)	(16.053)
BG on Q3	-0.017	-27.119**	0.030	-7.249
	(0.091)	(12.926)	(0.080)	(15.515)
More Successful on Q3	-0.009	-16.422	-0.041	-11.012
	(0.083)	(10.405)	(0.068)	(12.534)
BG on Q4	0.039	-10.621	0.072	7.437
	(0.035)	(37.735)	(0.043)	(32.575)
More Successful on Q4	0.050	-11.696	0.033	15.482
	(0.039)	(32.143)	(0.035)	(30.028)

Quartile refers to the quartile of the pre-intervention exam score, from Q1 (bottom-performing) to Q4 (top-performing). "Below 25% in Class" is a dummy variable for the student ranking below 5% in the first post-intervention exam. "Exam Ranking in the Grade" is the student's midterm exam ranking in Grade 7 or 8 at school. Control variables include baseline aggregate test score, homeroom teacher gender, a dummy variable for the homeroom teacher teaching a science subject, and a dummy variable indicating the student is in Grade 8. Each missing baseline variable except student gender is replaced by the median pre-intervention value and a dummy variable is included to capture this. All regressions include the school fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1.

Table D17: Heterogeneity in treatment impacts on students' aspiration for Zhong-Kao by baseline quartile

	Girls			Boys		
	(1) Aspire ranking Top 10%	(2) Aspire ranking Top 20%	(3) Aspire ranking Top 30%	(4) Aspire ranking Top 10%	(5) Aspire ranking Top 20%	(6) Aspire ranking Top 30%
T1: Very Successful	-0.069	0.018	-0.077	-0.193	-0.215*	-0.126
	(0.100)	(0.116)	(0.112)	(0.126)	(0.108)	(0.106)
T2: VS-General	0.149**	0.182**	0.032	-0.134	-0.065	0.047
	(0.064)	(0.085)	(0.102)	(0.125)	(0.128)	(0.124)
T3: Moderately Successful	-0.007	-0.005	-0.084	-0.016	-0.064	0.018
	(0.074)	(0.084)	(0.108)	(0.121)	(0.090)	(0.107)
T1 × Q2	0.026	-0.059	0.050	0.103	0.093	0.027
	(0.091)	(0.167)	(0.186)	(0.133)	(0.165)	(0.156)
T1 × Q3	-0.014	-0.079	0.044	0.151	0.239	0.158
	(0.145)	(0.187)	(0.154)	(0.172)	(0.164)	(0.160)
T1 × Q4	0.140	0.036	0.139	0.157	0.103	0.139
	(0.153)	(0.162)	(0.120)	(0.182)	(0.146)	(0.143)
T2 × Q2	-0.160*	-0.139	0.083	0.156	0.091	0.028
	(0.084)	(0.131)	(0.145)	(0.126)	(0.165)	(0.150)
T2 × Q3	-0.224	-0.218	-0.102	0.008	0.053	-0.001
	(0.143)	(0.172)	(0.162)	(0.171)	(0.192)	(0.178)
T2 × Q4	-0.090	-0.192	-0.068	0.157	-0.034	-0.066
	(0.125)	(0.151)	(0.135)	(0.176)	(0.170)	(0.151)
T3 × Q2	-0.088	-0.153	0.028	-0.045	-0.083	-0.052
	(0.080)	(0.132)	(0.169)	(0.129)	(0.141)	(0.138)
T3 × Q3	-0.093	-0.151	-0.094	-0.024	0.108	0.032
	(0.136)	(0.161)	(0.166)	(0.169)	(0.170)	(0.168)
T3 × Q4	0.082	0.124	0.128	0.054	0.089	0.011
	(0.135)	(0.130)	(0.135)	(0.165)	(0.131)	(0.150)
Q2: Lower-middle	0.083	0.227**	0.157	0.014	0.088	0.043
	(0.090)	(0.109)	(0.120)	(0.144)	(0.147)	(0.123)
Q3: Upper-middle	0.156	0.352**	0.257*	0.160	0.129	0.086
	(0.143)	(0.152)	(0.132)	(0.205)	(0.179)	(0.154)
Q4: Top-performing	0.264*	0.411**	0.275*	0.303	0.367**	0.170
	(0.152)	(0.164)	(0.151)	(0.217)	(0.179)	(0.150)
Constant	-0.034	0.040	0.312***	0.128	0.285*	0.448***
	(0.134)	(0.117)	(0.101)	(0.179)	(0.163)	(0.118)
Observations	820	820	820	853	853	853
R-squared	0.201	0.226	0.207	0.161	0.215	0.150
Overall treatment effects by groups:						
T1 on Q2	-0.043	-0.041	-0.027	-0.090	-0.122	-0.099
	(0.072)	(0.122)	(0.126)	(0.066)	(0.108)	(0.105)
T2 on Q2	-0.012	0.043	0.115	0.022	0.026	0.075
	(0.084)	(0.107)	(0.101)	(0.084)	(0.105)	(0.085)
T3 on Q2	-0.095	-0.158	-0.055	-0.062	-0.147	-0.035
	(0.080)	(0.110)	(0.112)	(0.076)	(0.101)	(0.090)
T1 on Q3	-0.083	-0.061	-0.033	-0.041	0.025	0.033
	(0.106)	(0.118)	(0.098)	(0.109)	(0.128)	(0.102)
T2 on Q3	-0.076	-0.036	-0.070	-0.125	-0.012	0.045
	(0.122)	(0.126)	(0.099)	(0.113)	(0.124)	(0.104)
T3 on Q3	-0.100	-0.156	-0.178*	-0.040	0.044	0.050
	(0.116)	(0.115)	(0.104)	(0.121)	(0.130)	(0.101)
T1 on Q4	0.071	0.055	0.062	-0.036	-0.111	0.013
	(0.096)	(0.089)	(0.054)	(0.113)	(0.078)	(0.074)
T2 on Q4	0.058	-0.010	-0.036	0.023	-0.099	-0.020
	(0.105)	(0.119)	(0.063)	(0.119)	(0.105)	(0.055)
T3 on Q4	0.075	0.119	0.044	0.037	0.025	0.029
	(0.093)	(0.074)	(0.058)	(0.109)	(0.084)	(0.077)
Effects of revealing different characteristics of role models, by groups:						
BG on Q1	-0.218**	-0.164	-0.109	-0.059	-0.150	-0.172*
	(0.083)	(0.109)	(0.115)	(0.087)	(0.128)	(0.098)
More Successful on Q1	-0.062	0.024	0.007	-0.176*	-0.151	-0.143*
	(0.092)	(0.112)	(0.121)	(0.092)	(0.096)	(0.078)
BG on Q2	-0.031	-0.084	-0.142	-0.112	-0.148	-0.174*
	(0.068)	(0.103)	(0.102)	(0.083)	(0.095)	(0.089)
More Successful on Q2	0.052	0.118	0.028	-0.028	0.025	-0.065
	(0.072)	(0.111)	(0.113)	(0.073)	(0.092)	(0.093)
BG on Q3	-0.007	-0.025	0.037	0.084	0.037	-0.013
	(0.078)	(0.122)	(0.098)	(0.087)	(0.087)	(0.079)
More Successful on Q3	0.017	0.095	0.145	-0.001	-0.019	-0.017
	(0.074)	(0.111)	(0.098)	(0.104)	(0.099)	(0.082)
BG on Q4	0.013	0.065	0.098	-0.059	-0.012	0.033
	(0.125)	(0.127)	(0.071)	(0.126)	(0.115)	(0.078)
More Successful on Q4	-0.004	-0.064	0.018	-0.073	-0.136	-0.016
	(0.119)	(0.085)	(0.065)	(0.125)	(0.106)	(0.097)

Quartile refers to the quartile of the pre-intervention exam score, from Q1 (bottom-performing) to Q4 (top-performing). "Aspire ranking Top 10%, 20%, or 30%" refers to student aspirations of ranking top 10%, 20%, or 30% in the Senior High School Entrance Examination (aka Zhong-Kao). Each missing baseline variable, except student gender, is replaced by the median pre-intervention value, and a dummy variable is included to capture this. Control variables include student gender, student baseline aggregate test scores, student baseline reported aspiration for Zhong-Kao, homeroom teacher gender, a dummy for homeroom teachers teaching a science subject, a dummy for homeroom teachers teaching Chinese literature, and a dummy variable indicating the student is in Grade 8. All regressions included the school fixed effects and grade fixed effects. Standard errors in parentheses are clustered at the unit of randomization (class level). *** p<0.01, ** p<0.05, * p<0.1.

Table D18: Pre-Analysis Plan Discrepancies

Pre-analysis plan	Modification	Location
Outcome variables		
Poor mental health index	PAP does not specify this index. Group the *Stress* and the *Depression* variables following the method in Anderson (2008). This paper considers the mental health variables as one of the primary outcomes of interest.	Main paper Tables 2 & 3, Appendix Tables D8, D9, D12, D13, & D15
Test scores of each subject	Only report the effects on math and overall test scores. Test scores of other subjects are omitted for space reasons.	Omitted
Test scores of each subject	Only report the effects on math and overall test scores. Test scores of other subjects are omitted for space reasons.	Omitted
Results of Senior High School Entrance Exam	PAP does not include the analysis of these long-term test scores. The analysis follows the same regression as that used to analyze the primary outcomes of interest. These related variables will update later after the Grade 7 students complete their Entrance Exam.	Appendix C
Others		
Regression equation	The grade FE is replaced by a dummy variable for the student being in Grade 8. *They are equivalent statistically because my sample contains only two grades. See Eq.5 for details.*	Results remain.
Main regression results	Add sharpened q-values following the method of Benjamini et al. (2006).	Main paper Tables 2 & 3.
Heterogeneous treatment effects by baseline ability	Instead of checking the heterogeneity by the *below_median* dummy, I divide the students into four groups based on their baseline total test scores.	Figure 5, Appendix Tables D13, D14, &D15.
Heterogeneous treatment effects by multiple baseline variables	This is an additional analysis of heterogeneity using the method of causal forest.	Appendix B and Appendix Table D12
Additional Robustness checks	1. Plot CDFs of the test scores and perform two-sample Kolmogorov-Smirnov tests.	Appendix Figure D1
	2. Scope of the spillover effects.	Appendix Table D7
	3. Permutation test for the treatment effects.	Appendix Table D8
	4. Treatment effects without control variables.	Appendix Table D9

E Proof of the Model Equilibrium Condition

The student's optimization problem is

$$\max_{e} E[U] = U_1(y_1, e) + \beta E[U_2(z)] = [\bar{u}_1 - c(e)] + \beta\big\{E[u_2(y_2)] - E[\psi(\alpha - y_2)]\big\}, \tag{E1}$$

where $y_2 = y_1 + g(e) + \varepsilon$ is the test score in Period 2. $g(e)$ indicates additional score from effort e, satisfying $g'(e) > 0$ and $g''(e) < 0$. An error term $\varepsilon \sim N(0, \sigma^2)$ presents the uncertainty. So y_2 also follows a normal distribution, such that $y_2 \sim N(y_1 + g(e), \sigma^2)$.

I assume that $u_2(s) = s$. Thus,

$$E[u_2(y_2)] = E[y_2] = y_1 + g(e) + E[\varepsilon] = y_1 + g(e). \tag{E2}$$

I denote the CDF and PDF of the standard normal distribution as $\Phi(\cdot)$ and $\phi(\cdot)$, respectively. Then, the expectation of the frustration term can be written as

$$\begin{aligned} E[\psi(\alpha - y_2)] &= \lambda \int_{-\infty}^{\alpha} (\alpha - y_2) f(y_2) dy_2 \\ &= \lambda[(\alpha - y_1 - g(e))\Phi(z) + \sigma\phi(z)], \end{aligned} \tag{E3}$$

where $z = \frac{\alpha - y_1 - g(e)}{\sigma} \sim N(0, 1)$. Thus,

$$\alpha - y_1 - g(e) = \sigma z, \text{ and } \frac{dz}{de} = -\frac{g'(e)}{\sigma}. \tag{E4}$$

The PDF of z, $\phi(z)$ follows $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$, therefore, the derivative of $\phi(z)$ is

$$\frac{d}{de}\phi(z) = \frac{1}{\sqrt{2\pi}} \frac{d}{de}\big[e^{-\frac{z^2}{2}}\big] = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} (-z) \frac{dz}{de} = z\phi(z)\frac{g'(e)}{\sigma}. \tag{E5}$$

The derivative of Equation E1 is

$$\begin{aligned}
\frac{dE[U]}{de} &= -c'(e) + \beta\big\{g'(e) - \lambda\frac{d}{de}[(\alpha - y_1 - g(e))\Phi(z) + \sigma\phi(z)]\big\} \\
&= -c'(e) + \beta g'(e) - \beta\lambda\big\{\frac{d}{de}(\alpha - y_1 - g(e))\Phi(z) + \sigma\frac{d}{de}\phi(z)\big\} \\
&\overset{(E5)}{=} -c'(e) + \beta g'(e) - \beta\lambda\big\{\frac{d}{de}(\alpha - y_1 - g(e))\Phi(z) + \sigma z\phi(z)\frac{g'(e)}{\sigma}\big\} \\
&= -c'(e) + \beta g'(e) - \beta\lambda\big\{\Phi(z)\frac{d}{de}(\alpha - y_1 - g(e)) + (\alpha - y_1 - g(e))\phi(z)\frac{dz}{de} + z\phi(z)g'(e)\big\} \\
&\overset{(E4)}{=} -c'(e) + \beta g'(e) - \beta\lambda\big\{\Phi(z)(-g'(e)) + \sigma z\phi(z)(-\frac{g'(e)}{\sigma}) + z\phi(z)g'(e)\big\} \\
&= -c'(e) + \beta g'(e) - \beta\lambda\big\{\Phi(z)(-g'(e)) - z\phi(z)g'(e) + z\phi(z)g'(e)\big\} \\
&= -c'(e) + \beta g'(e) + \beta\lambda\Phi(z)g'(e)
\end{aligned} \tag{E6}$$

The first-order condition satisfies

$$\frac{dE[U]}{de} = -c'(e) + \beta g'(e) + \beta\lambda\Phi(z)g'(e) = 0.$$

Therefore,

$$\beta[g'(e) + \lambda\Phi(z)g'(e)] = c'(e),$$

where $\Phi(z) = \Phi(\frac{\alpha - y_1 - g(e)}{\sigma}) = Pr[\frac{y_2 - y_1 - g(e)}{\sigma} < \frac{\alpha - y_1 - g(e)}{\sigma}] = Pr[y_2 < \alpha]$ indicates the probability of not achieving the aspiration α. #