Breda, Thomas; Grenet, Julien; Monnet, Marion; Van Effenterre, Clémentine

Working Paper

Do Female Role Models Reduce the Gender Gap in Science? Evidence from French High Schools

IZA Discussion Papers, No. 13163

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Breda, Thomas; Grenet, Julien; Monnet, Marion; Van Effenterre, Clémentine (2020) : Do Female Role Models Reduce the Gender Gap in Science? Evidence from French High Schools, IZA Discussion Papers, No. 13163, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/216475

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13163

Do Female Role Models Reduce the Gender Gap in Science? Evidence from French High Schools

Thomas Breda
Julien Grenet
Marion Monnet
Clémentine Van Effenterre

APRIL 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13163

Do Female Role Models Reduce the Gender Gap in Science? Evidence from French High Schools

Thomas Breda
CNRS, Paris School of Economics and IZA

Julien Grenet
CNRS and Paris School of Economics

Marion Monnet
Paris School of Economics

Clémentine Van Effenterre
University of Toronto

APRIL 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Do Female Role Models Reduce the Gender Gap in Science? Evidence from French High Schools*

This paper, based on a large-scale field experiment, tests whether a one-hour exposure to external female role models with a background in science affects students' perceptions and choice of field of study. Using a random assignment of classroom interventions carried out by 56 female scientists among 20,000 high school students in the Paris Region, we provide the first evidence of the positive impact of external female role models on student enrollment in STEM fields. We show that the interventions increased the share of Grade 12 girls enrolling in selective (male-dominated) STEM programs in higher education, from 11 to 14.5 percent. These effects are driven by high-achieving girls in mathematics. We find limited effects on boys' educational choices in Grade 12, and no effect for students in Grade 10. Evidence from survey data shows that the program raised students' interest in science-related careers and slightly improved their math self-concept. It sharply reduced the prevalence of stereotypes associated with jobs in science and gender differences in abilities, but it made the underrepresentation of women in science more salient. Using machine learning methods, we leverage the diversity of role model profiles to document substantial heterogeneity in the effectiveness of role models and shed light on the channels through which they can influence female students' choice of study. Results suggest that emphasis on the gender theme is less important to the effectiveness of this type of intervention than the ability of role models to convey a positive and more inclusive image of STEM careers.

JEL Classification: C93, I24, J16

Keywords: role models, gender gap, STEM, stereotypes, choice of studies

Corresponding author:
Clémentine Van Effenterre
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138
USA

E-mail: clementine_van_effenterre@hks.harvard.edu

* We are grateful to the staff at the L'Oréal Foundation, especially Diane Baras, Aude Desanges, Margaret Johnston-Clarke, Salima Maloufi-Talhi, David McDonald, and Elisa Simonpietri for their continued support to this project. We also thank the staff at the French Ministry of Education (Ministère de l'Éducation Nationale, Direction de l'Évaluation, de la Prospective et de la Performance) and at the Rectorats of Créteil, Paris, and Versailles for their invaluable assistance in collecting the data. This paper greatly benefited from discussions and helpful comments from Marcella Alsan, Iris Bohnet, Scott Carrell, Clément de Chaisemartin, Bruno Crépon, Esther Duflo, Lena Edlund, Ruth Fortmann, Pauline Givord, Laurent Gobillon, Marc Gurgand, Élise Huillery, Philip Ketz, Sandra McNally, Amanda Pallais, and Liam Wren-Lewis. We thank participants at the ASSA/AEA Annual Meeting in Atlanta, CEPR/IZA Annual Symposium in Labor Economics 2018 in Paris, EALE 2018 in Lyon, EEA-ESEM 2018 in Cologne, Gender Economics Workshop 2018 in Berlin, Gender and Tech Conference at Harvard, IWAEE Conference in Catanzaro, and Journées LAGV 2018 in Aix-Marseille. We also thank seminar participants at Bristol, DEPP, LSE, Harvard Kennedy School, HEC Lausanne, Maastrict SBE, OECD EDU Forum, PSE, Stockholm University SOFI, and Université Paris 8. We are grateful to the Institut des politiques publiques (IPP) for continuous support and to Sophie Cottet for her assistance in contacting schools. Financial support for this study was received from the Fondation L'Oréal and from the Institut des Politiques Publiques. The project received IRB approval at J-PAL Europe and was registered in the AEA RCT Registry with ID AEARCTR-0000903.

Introduction

Women's increasing participation in science and engineering in the U.S. has leveled off in the past decade (National Science Foundation, 2017). This trend, which is common to almost all OECD countries, is a source of concern for two main reasons. First, it exacerbates gender inequality in the labor market, as Science, Technology, Engineering, and Mathematics (STEM) occupations offer higher average salaries (Brown and Corcoran, 1997; Black et al., 2008; Blau and Kahn, 2017) and show a smaller gender wage gap (Beede et al., 2011). Second, in a context of heightened concern over a shortage of STEM workers in the advanced economies (Carnevale et al., 2011; Xue and Larson, 2015), this trend is likely to represent a worsening loss of talent that could reduce aggregate productivity (Weinberger, 1999; Hoogendoorn et al., 2013).

The underrepresentation of women in these traditionally male-dominated fields can also constitute a self-fulfilling prophecy for subsequent generations, as girls have little opportunity to interact with women working in these fields and who could inspire them. Exposing female students to successful or admirable women scientists could help to break this vicious circle. These "role models" potentially extend female students' possibility set, raise their aspirations, alleviate stereotype threat, and provide relevant information.

A large body of work has established that female science teachers and professors can serve as role models and that they help improve female students' academic achievement (Dee, 2007; Hoffmann and Oreopoulos, 2009; Eble and Hu, 2017; Lim and Meer, 2017), increase their probability of enrolling in STEM majors (Bettinger and Long, 2005; Carrell et al., 2010; Lim and Meer, 2019), and influence their occupational choices (Kofoed and McGovney, 2019).[1] However, these effects are not easy to interpret, in that they could be driven by gender differences in teaching practices and behavior in class rather than by female teachers acting as role models. The positive influence of female instructors on female students might derive from greater encouragement received over an entire academic year (Lavy and Sand, 2018; Terrier, forthcoming) or from implicit biases, which have been shown to be more pronounced among male than among female teachers (Carlana, 2019). The policy implications of the studies on teacher-student gender interactions are also unclear, since the scarcity of female instructors in STEM fields restricts the scope for increasing the exposure of girls to this group, at least in the short run.

[1]Seminal papers on the impact of teacher-student gender interactions include Canes and Rosen (1995), Rothstein (1995), and Neumark and Gardecki (1998). More recent studies have investigated the impact of teachers' characteristics on students' future careers in developing countries (Paredes, 2014; Muralidharan and Sheth, 2016) and the effects of other types of gender interaction, e.g., between students and advisors (Canaan and Mouganie, 2019) or between children and doctors (Riise et al., 2019).

Our paper overcomes these limitations by providing direct empirical evidence on the impact of non-teaching role models on student outcomes. We use a large-scale randomized experiment to test whether a one-hour exposure to external female scientists acting as role models can change high school girls' perceptions of science careers, and ultimately steer them towards STEM studies. We leverage rich administrative and survey data on nearly 20,000 high school students and 56 female scientists to investigate the mechanisms underlying the effect of role models on students' choice of field of study, both in high school and at college entry.

The program we evaluate is called "For Girls and Science" (*Pour les Filles et la Science*) and was launched in 2014 by the L'Oréal Foundation—the corporate foundation of the world's leading cosmetics manufacturer—to encourage girls to explore STEM career paths. It consists of one-hour in-class interventions by women with two very distinct profiles: half are young scientists (either Ph.D. candidates or postdoctoral researchers) who were awarded the L'Oréal-UNESCO "For Women in Science" Fellowship; the others are young professionals privately employed as scientists in the Research and Innovation division of the L'Oréal group. In the main part of the intervention, the role models share their experience and career path with the students. They also provide consistent information on science-related careers in general and more specifically on the underrepresentation of women, using two short videos and a set of customizable slides.

The evaluation was conducted during the 2015/16 academic year in 98 of the 489 public and private high schools located in the Paris region. It involved 19,451 students from 416 classes in Grade 10 and 185 classes in Grade 12 (science track). Half of the classes were randomly assigned to be visited by one of the 56 role model participants. Important, irreversible educational choices are made by students at the end of both Grade 10 and Grade 12, allowing us to study the effect of the program on STEM enrollment outcomes in the following year, i.e., 2016/17.

The first contribution of our paper is to provide credible evidence of the impact of external female role models on STEM enrollment decisions in high school and at entry in higher education. In our empirical setting, we show that the role models' interventions led to a significant increase in the share of girls enrolling in STEM fields, but only in the educational tracks where they are strongly underrepresented. In Grade 10, the classroom visits had no detectable impact on boys' and girls' probability of enrolling in the science track in Grade 11, where girls are only slightly underrepresented (47 percent of students). By contrast, the program induced a significant 2.4-percentage-point increase in STEM undergraduate enrollment among girls in Grade 12, or an increase of 8 percent over the baseline rate of 29 percent, while the effect for boys was negligible. This positive impact on female STEM enrollment is driven by female students shifting to selective STEM programs, which lead to the most prestigious graduate schools, and

male-dominated STEM programs (math, physics, computer science, and engineering). The probability of enrolling in such programs after high school increased by 20 to 30 percent among female students, which corresponds to one girl in every two Grade 12 classes switching to either type of program at college entry. These effects, which are robust to a wide range of specification tests, can be expected to improve the future earnings of the target group, as selective and male-dominated STEM programs feature not only a low share of female students (around 30 percent) but also substantial wage premia relative to non-STEM and female-dominated STEM programs, in the order of 10 to 25 percent at entry level.

Our paper's second main contribution is to explore the channels through which role models affect students' choice of study. Specifically, we analyze a wide range of potential mediators, from perceptions of science-related jobs and gender role in science to academic performance, self-concept, and career aspirations. To analyze these outcomes, we conducted a post-treatment student survey consisting of an eight-page questionnaire administered in class one to six months after the classroom interventions. We also collected administrative data on high school graduation exams (*Baccalauréat*) at the end of Grade 12.

Our results show that the program had no significant effect on students' self-reported taste for science subjects or their academic performance, and only slightly increased their math self-concept. By contrast, the role model interventions significantly improved students' perceptions of science-related jobs at both grade levels, with no indication of declining effects over a period of up to six months. For girls in Grade 12, the program also helped mitigate some of the masculine stereotypes typically associated with STEM occupations (such as being hard to reconcile with family life) and heightened the perception that these jobs pay better. Consistently, we find that unlike those of boys, girls' aspirations for science-related careers increased significantly in Grade 12.

One of the most interesting—and least expected—findings concerns the effects on students' perceptions of gender roles in science. The classroom interventions not only were effective in debiasing students' beliefs about gender differences in math aptitude, they also raised awareness of the underrepresentation of women in science. The combination of these two effects triggered an unintended ex-post rationalization by students of the gender imbalance in scientific fields and occupations, making them more likely to think that women dislike science and that they face discrimination in science-related jobs. Explicitly addressing the gender gap in science thus appears to have generated more ambiguous perceptions among students than the program's gender-neutral messages about jobs and careers.

Towards a better understanding of the mechanisms underlying the effects of role models on

student behavior, we implement the generic approach recently developed by Chernozhukov et al. (2018) to analyze patterns of heterogeneity using machine learning methods. The results indicate a high degree of heterogeneity in treatment effects according to student performance and role model background. We find that the significant impact on female STEM enrollment after Grade 12 is driven by high-achieving girls in math. Highlighting the importance of the role model component of the program, we show, further, that the young professionals employed by the sponsoring firm had a significantly greater effect on girls' probability of enrolling in selective STEM programs than the young researchers. While the two groups were equally effective in debunking the stereotype on gender differences in math aptitude, there is clear evidence that those with a professional background were better able to improve girls' perceptions of science-related jobs and raise their aspirations for such careers. Conversely, they were less likely to reinforce students' belief that women are underrepresented in science.

Showing the critical role played by the person who bears the message, these comparisons demonstrate that role model interventions are not reducible to the provision of standard information content and that female role models are not interchangeable. They also suggest that the gender debiasing component of the classroom interventions, which emphasized men and women's equal predisposition for science, cannot explain, alone, the program's effect on girls' STEM enrollment outcomes.

To confirm these insights, we carry out a more systematic investigation of mechanisms, exploiting the rich variation in treatment effects across students' observable characteristics and the profiles of the 56 role models. Building on Chernozhukov et al. (2018), we develop a new method to estimate the correlations between the individual-level treatment effects on STEM enrollment outcomes and the treatment effects on potential mediators. The results show that the role models who had the greatest effect on female enrollment in selective STEM programs are those who most improved girls' perceptions of science-related careers without reinforcing the perception that women are underrepresented in science. By contrast, we find that the role models' ability to steer girls towards selective STEM programs is essentially uncorrelated with their effects on students' perceptions of gender differences in aptitude for science.

Overall, our study offers strong evidence that short-term exposure to non-teaching female role models with a background in science can significantly increase female participation in the most selective and male-dominated STEM fields of study at college entry—these fields being important contributors to the gender pay gap. Our exploration of mechanisms provides consistent evidence that the emphasis on gender issues is less important to the effectiveness of such interventions than the ability of role models to project a positive and inclusive image of

science-related careers, thus embodying an attractive, attainable path to them.

Related literature. Our paper relates to three strands of the literature. First, it adds to the extensive body of research on the origins of the gender gap in STEM. Gender differences in aptitude are unlikely to explain this gap (Hyde, 2005; Spelke, 2005) and some scholars argue that direct discrimination is no longer a major determinant (Ceci et al., 2014). Increasingly, attention has focused on understanding gender differences in the educational choices made by equally able students. Parents, schools, and teachers are often said to convey stereotypes and social norms that influence these choices and so, in the long run, contribute to maintaining strong gender segregation across school majors.[2] Our paper shows that a brief exposure to external (non-teaching) female scientists can counteract these influences, significantly changing high school girls' perceptions of science careers and affecting their choice of field of study. These interventions are particularly policy-relevant, as they can easily be scaled up.

Second, the paper contributes to a thin but growing literature on the effect of non-teaching role models on educational outcomes. While most of the economics literature on role model effects has focused on the influence of gender on teacher-student interactions, a handful of studies in psychology and economics have analyzed the effects of non-teaching role model interventions using lab experiments (Lockwood and Kunda, 1997; Dasgupta and Asgari, 2004; Cheryan et al., 2011; Betz and Sekaquaptewa, 2012; O'Brien et al., 2016) or in the field (Nguyen, 2008; Beaman et al., 2012; Burgess, 2016; Del Carpio and Guadalupe, 2018; Ashraf et al., forthcoming; Riise et al., 2019; Porter and Serra, forthcoming). For instance, Beaman et al. (2012) find that exposure to women in leadership positions in India has a positive impact on girls' educational attainment and on parents' career aspirations for daughters (but not sons). More closely related to our paper is the study by Porter and Serra (forthcoming), which documents a positive impact of two female role models who were carefully selected among the economics alumni of Southern Methodist University in the U.S. on the likelihood of female students majoring in economics. Our study is the first to focus on the ability of light-touch role model interventions on a large scale to steer female students towards STEM career paths. In contrast to previous studies, which typically involve only a small number of female role models, the combination of a rich survey of students' perceptions and a large number of role model participants allows analyzing the mechanisms in greater depth.

Finally, our study contributes to the rapidly expanding literature on the use of machine

[2]These social pressures and gender stereotypes do not necessarily translate into explicit discrimination (Ceci and Williams, 2011; Breda and Ly, 2015; Breda and Hillion, 2016) but rather seem to be mostly internalized and thereby influence academic self-perception (Correll, 2001; Ehrlinger and Dunning, 2003) and behavior in competitive environments (Gneezy et al., 2003; Niederle and Vesterlund, 2007, 2010; Buser et al., 2014), fostering an environment conducive to self-censorship (Babcock and Laschever, 2012; Leibbrandt and List, 2015).

learning to analyze heterogeneity in individual treatment effects (Athey and Imbens, 2016, 2017; Mullainathan and Spiess, 2017; Wager and Athey, 2018). To the best of our knowledge, this paper is among the first to implement the generic approach proposed by Chernozhukov et al. (2018) to conduct inference on key features of heterogeneous effects. We also devise an empirical strategy to estimate the correlations between student-level treatment effects on different outcomes conditional on observable exogenous characteristics. This constitutes a methodological contribution that may be used to investigate channels of influence in randomized controlled trials.

Organization. The remainder of the paper is organized as follows. Section 1 provides some institutional background on the French educational system and the gender gap in STEM fields. Section 2 describes the program and the experimental design. Section 3 presents the data and empirical strategy. Section 4 analyzes the effects of role model interventions on student perceptions, self-concept, and educational outcomes. Section 5 extends the analysis to the persistence of effects, the role of the timing of interventions, and potential spillovers. Section 6 discusses potential mechanisms, and Section 7 concludes.

1 Institutional Background

1.1 Structure of the French Education System

In France, education is compulsory from the age of 6 to the age of 16, with the academic year running from September to June. The school system consists of five years of elementary education (Grades 1 to 5) and eight years of secondary education, divided into four years of middle school (*collège*, Grades 6 to 9) and three of high school (*lycée*, Grades 10 to 12). Students complete high school with the *Baccalauréat* national exam, which they must pass for admission to higher education.

High school tracks. The tracking of students occurs at two critical stages (see Figure 1). At the end of middle school, about two-thirds of students are admitted to general and technological upper secondary education (*Seconde générale et technologique*) and the remaining third are tracked into vocational schools (*Seconde professionnelle*). After the first year of high school (Grade 10), the general and technological track is further split: approximately 80 percent of the students are directed to the general *Baccalauréat* program for the last two years of high school (Grades 11 and 12) and 20 percent are directed towards a technological *Baccalauréat*.

In the Spring term of Grade 10, the students who have been allowed to pursue the general

track are required to choose among three sub-tracks in Grade 11: Science (*Première S*), Humanities (*Première L*), and Social sciences (*Première ES*). This is an important choice, given that the curriculum and high school examinations are specific to each *Baccalauréat* track and thus directly impact on students' educational opportunities and career prospects.[3] It is almost impossible, for instance, for a student to be admitted to engineering or medical undergraduate programs without a *Baccalauréat* in science. Students directed to the technological track after Grade 10 are also required to choose among eight possible STEM and non-STEM sub-tracks, which will affect their choice of field of study in higher education.[4]

College entry. In the Spring term of Grade 12, students in their final year of high school apply for admission to higher education programs through a centralized online admission platform. The programs to which students can apply fall into two broad categories, each accounting for about half of first-year undergraduate enrollment: (i) non-selective undergraduate university programs (*Licence*), which are open to all students who hold the *Baccalauréat*; and (ii) selective programs. Both types of program offer specializations in STEM and non-STEM fields. Non-selective programs cannot select students based on their academic record,[5] while selective programs can admit students based on their academic achievement. Among selective programs, the most prestigious are the two-year *Classes préparatoires aux Grandes Écoles* (CPGE), which prepare students to take the national entry exams to elite graduate schools (*Grandes Écoles*). These programs are specialized either in science, economics and business or in humanities. Within the science CPGE programs, the main fields of specialization are mathematics and physics (MPSI), physics and chemistry (PCSI), and biology/geoscience (BCPST). The other selective undergraduate programs (*Section de technicien supérieur* or STS) are mostly targeted to students holding a vocational or technological *Baccalauréat* and prepare for technical/vocational bachelor's degrees.

[3]STEM-specific tracking in upper secondary education is not unique to France. In many countries, high school students undertaking general academically oriented programs are channelled through discipline-based tracks (Marginson et al., 2013).

[4]The technological track in high school includes two STEM sub-tracks (STI2D: *Sciences et technologies de l'industrie et du développement durable*; and STL: *Sciences et technologies de laboratoire*) and six non-STEM sub-tracks (S2TMD: *Sciences et techniques du théâtre, de la musique et de la danse*; ST2S: *Sciences et technologies de la santé et du social*; STAV: *Sciences et technologies de l'agronomie et du vivant*; STD2A: *Sciences et technologies du design et des arts appliqués*; STHR: *Sciences et technologies de l'hôtellerie et de la restauration*; and STMG: *Sciences et technologies du management et de la gestion*).

[5]When the number of applications to a non-selective program exceeds the number of seats available, students from the university's academic region (*académie*) are given priority over other students and, within the group of local applicants, students who have ranked the program at a higher position in their rank-ordered list are given priority; remaining ties are broken at random.

1.2 Female Underrepresentation in STEM

In France, the share of female students in STEM-oriented studies starts to decline after Grade 10 and drops sharply at entry into higher education (see Figure A1 in the Appendix). While 54 percent of the students in the general and technological track in Grade 10 are girls, the share falls to 47 percent in the general science track (Grades 11 and 12) and further to 30 percent in the first year of higher education.[6] Female underrepresentation in STEM fields of study is more pronounced in the selective undergraduate programs (shares of 18 percent in STS and 30 percent in CPGE) than in the non-selective programs (35 percent). These proportions, which are computed from administrative data for 2016/17, are almost identical to those of a decade earlier. Within STEM fields of study, female students tend to specialize in earth and life sciences rather than mathematics, physics, or computer science (see Appendix Figure A2).

The underrepresentation of women in STEM fields accounts for a good part of the gender pay gap among college graduates in France. Using a variety of administrative and survey data sources, we show in Appendix B that across all majors, male graduates who obtained a master's degree in 2015 or 2016 earn a median gross annual starting salary of 32,122 euros, compared to 28,411 euros for female graduates. This amounts to an overall gap of 3,711 euros per year, or 11.6 percent of men's pay. Using standard decomposition methods, we find that the underrepresentation of female students in STEM accounts for approximately 25 percent of this gap. Almost half of the gender pay gap within STEM can be ascribed to the fact that female graduates are less likely than males to be enrolled in the selective and male-dominated fields, which lead to the best-paying degrees. These figures strongly suggest that in the French context, increasing the share of female students in STEM—especially in selective and male-dominated programs—would narrow the gender pay gap substantially.

2 Program and Experimental Design

2.1 The Program

The program "For Girls and Science" (FGiS) is an awareness campaign launched in 2014 by the L'Oréal Foundation to encourage girls to explore STEM career paths. It consists in one-hour one-off classroom interventions by female role models with a background in science. It originated in France and was later extended to other countries, including Canada, Italy and

[6] At the high school level, the gender imbalance in STEM is more severe in the technological than in the general science track; the female share is as low as 17 percent in the two STEM-oriented technological sub-tracks (STI2D and STL).

New Zealand. The interventions, which take place in the presence of all students in the class, including boys, are carried out by female role models of two distinct types: (i) Ph.D. candidates or post-doctoral researchers who have been awarded a fellowship by the L'Oréal Foundation (the L'Oréal-UNESCO "For Women in Science" Fellowship) and who participate in the program as part of their contract;[7] and (ii) young professionals employed as scientists in the Research and Innovation division of the L'Oréal group who volunteer for the program.

Structure and content of the interventions. The classroom interventions last one hour and are divided into four main sequences. The presentation begins with a set of customizable slides that highlight two facts: (1) the labor market is marked by high demand for STEM skills, and there is a shortage of graduates in the relevant fields of study; and (2) women are underrepresented in STEM careers. These two messages are illustrated with examples of career prospects in humanities versus science, emphasizing differences in employment rates, average earnings, and the prevalence of gender segregation in high-wage occupations. The slides further stress the contribution of female underrepresentation in STEM to the gender pay gap.

The second sequence kicks off with two three-minute videos designed to illustrate and deconstruct stereotypes about science-related careers and gender roles in science.[8] The first video, entitled "Science, Beliefs or Reality?," uses interviews with high school students to debunk myths about careers in science (e.g., jobs in science are more challenging, they necessarily require long studies), stereotypes about scientists (e.g., they are introverted, lonely), and gender differences in science aptitude (e.g., women are naturally less talented in math). The second video, entitled "Are we all Equal in Science?," describes the common gender stereotypes about aptitude for science while providing information on brain plasticity and on how interactions and the social environment shape men's and women's abilities and tastes. This sequence aims at stimulating class discussion based on students' reactions to the videos.

The third sequence centers on the female role model's own experience as a woman with a background in science and consists of an interactive question-and-answer session with the students. Topics addressed during this discussion include the role model's typical day at work, her everyday interactions with co-workers, how much she earns, and work-family balance. Consistent with the program's emphasis on the "role model" dimension, this sequence was intended to be the longest and most important part of the intervention. In order to convey this objective to the role models, a full-day training was organized to help them share their

[7] In 2015/16, the individual L'Oréal-UNESCO "For Women in Science" fellowships amounted to 15,000 euros for Ph.D. candidates and to 20,000 euros for postdoctoral researchers.

[8] Screenshots of the two videos shown during the classroom interventions are displayed in Appendix Figure C3.

experience with the students. The training also included a workshop on the underrepresentation of women in science and a practice session aimed at enhancing oral communication skills.

The intervention concludes with an overview of the diversity of STEM studies and careers, illustrated by concrete examples such as jobs in graphic design, environmental engineering, and computer science.

2.2 Experimental Design

Selection of schools and classes. The evaluation was conducted in the three education districts (*académies*) of the Paris region (Paris, Créteil, and Versailles) during the 2015/16 academic year, the program's second year of existence. Créteil and Versailles are the two largest education districts in France and the three districts combined include 318,000 high school students in the general and technological track, or 20 percent of all French high school enrollment.

Figure 2 provides a detailed timeline of the evaluation. In the spring of 2015, the French Ministry for Education agreed to support a randomized evaluation of the program and designated one representative for each district as intermediary between the schools and the evaluation team. In June, official letters informed high school principals that they were likely to be contacted to take part in the evaluation. All public and private high schools with at least four classes in Grade 10 and two in Grade 12 (science track) were contacted by our team between September and December 2015, accounting for 349 of the 489 high schools operating in the three districts. Of these schools, 98 agreed to take part in the experiment, representing 28 percent of Grade 10 enrollment and 29 percent of Grade 12 (science track) enrollment in the three districts combined.[9] The overall sample, which consists of 19,451 students (13,700 in Grade 10 and 5,751 in Grade 12), closely resembles the relevant student population in the Paris region, both in social composition and in average academic performance (see Appendix Table F3).

Randomization. In the fall of the 2015/16 school year, the principals were invited to select at least six classes—four or more in Grade 10 and two or more in Grade 12 (science track)—and to indicate a preferred time slot and day for the interventions.[10] In each school, half of the classes selected by the principal (up to the nearest integer) were randomly assigned to the treatment group (302 classes in total) and the other half to the control group (299 classes). Table 1 indicates that the random assignment successfully balanced the characteristics of students in

[9]The location of the participating schools is shown in Appendix Figure C4.

[10]In the vast majority of schools, principals selected exactly four Grade 10 and two Grade 12 classes.

the treatment and control groups.[11]

Role models. The experiment involved 56 female role models, of whom 35 were L'Oréal employees and 21 were Ph.D. candidates or post-doctoral researchers. Table 2 provides summary statistics on their characteristics. The researchers tend to be younger (30 vs. 36 years of age on average) and are less often of foreign nationality (10 vs. 17 percent). Although both types have very high levels of educational attainment, 39 percent having graduated from a *Grande École*, the researchers are more likely than the professionals to to hold (or prepare for) a Ph.D. (100 vs. 38 percent) and to hold a degree in math, physics and engineering (38 vs. 14 percent). They are also less likely to have children (19 vs. 58 percent) and to have been involved in the program in the previous year (19 vs. 29 percent). Although we could not collect direct information on earnings, for reasons of confidentiality, we estimate the annual gross wages of the young professionals to be between 45,000 and 65,000 euros, compared to 22,000–50,000 euros for the researchers.[12] On average, each role model carried out five classroom interventions in two different high schools.

Classroom interventions. The classroom visits took place between November 17, 2015, and March 3, 2016.[13] The role models were asked to select two or three schools in which to carry out an average of three classroom visits per school—in most cases, two in Grade 10 and one in Grade 12. They were not assigned to the schools randomly but registered for the visits and time slots using an online system on a first-come, first-served basis.[14]

[11] Additional balancing checks by gender are presented in Appendix Tables F5 and F6.

[12] The typical range of annual gross salaries for the professionals is estimated based on information we obtained from the Human Resources Department of the L'Oréal Group, which indicates that 80 percent of the 30 to 39 year-old female executives employed by L'Oréal earn an annual gross wage between 45,000 and 65,000 euros. For researchers, the range of annual gross wages can be estimated to be between 22,000 euros, which corresponds to the basic gross annual salary for Ph.D. contracts at French public higher education institutions, and 50,000 euros, which corresponds to the high end of the range of annual salaries earned by postdocs in France.

[13] 17 percent of the visits in November, 26 percent in December, 39 percent in January, 17 percent in February, and 1 percent in March.

[14] Randomly assigning the role models to the schools was not a feasible option, since most were participating on a voluntary basis and during regular working hours. The non-random assignment does not represent a threat to identification, however, as the random assignment of classes to the treatment and control groups was stratified by school. Since participating schools were added to the schedule only gradually, multiple registration sessions were organized to match role model participants with the schools. All role models were contacted four times to complete the schedule, on October 21, November 24, December 7, 2015, and on February 3, 2016, thus limiting their ability to select the schools they would visit.

3 Data and Empirical Strategy

3.1 Data

To evaluate the program's effects on student perceptions and educational choices, we combine three main data sources: (i) a post-intervention survey of role models; (ii) a post-intervention survey of students; and (iii) student-level administrative data. Translated versions of the two surveys are provided in Appendix D.

Role model survey. After each visit to a school, which typically involved three consecutive classroom interventions, the role models were invited to complete an online survey. Besides collecting general feedback, this survey served to monitor compliance with random assignment, asking them to identify each of the classes they visited. Summary statistics are reported in Appendix Table F4. The interventions almost always (89 percent) took place in the presence of the teacher and sometimes (35 percent) of another adult. The role models reported organizational problems for only 16 percent of the visits (e.g. the intervention started late, the slides could not be shown). When asked about their overall perception of each of their classroom interventions, 93 percent said they went "well" (37 percent) or "very well" (56 percent). Students were generally perceived to be responsive to the key messages.

Student survey. We conducted a paper-and-pencil student survey in all participating classes one to six months after the classroom visits, between January and May 2016. Each questionnaire was assigned a unique identifier so that it could be linked with student-level administrative data. The survey was designed to collect a rich set of information on students' tastes, personality traits, choices and stereotypes, and was administered in exam conditions under the supervision of a teacher. It was presented as a general survey on students' attitudes about science and science-related careers so as to minimize the risk that students would associate it with the FGiS program. It was eight pages long and took about half an hour to complete.

The survey items investigated the effects of classroom interventions on students' perceptions and self-concept along five dimensions: (i) general perceptions of science-related careers; (ii) perceptions of gender roles in science; (iii) taste for science subjects; (iv) math self-concept; and (v) science-related career aspirations. When conceptually related, the survey items were combined to construct a synthetic index for each dimension using standardized z-score scales. Section 4 describes the specific items that we used for each dimension of interest.[15]

[15]To mitigate potential order bias, the order of several of the response items (e.g., math/French, man/woman)

On the last page of the questionnaire, the students in the treatment group were asked whether they had discussed the classroom intervention with their classmates, with schoolmates from other classes, or with friends outside of school, as a way of assessing possible spillover effects. Students in the control group received a slightly different version of this final section, asking whether they had heard of classroom visits by male or female scientists in other classes, with no explicit mention of the FGiS program.

As shown in Appendix Table F8, the survey response rates were high both in Grade 10 (88 percent of students) and in Grade 12 (91 percent). They were slightly higher among Grade 10 students in the treatment than in the control group (by 2.6 percentage points). Despite this small difference in response rates, Table F9 in the Appendix shows that the characteristics of survey respondents in Grade 10 are generally balanced between the treatment and control groups. The opposite is found in Grade 12: the survey response rates are similar in the two groups, but the respondents' characteristics exhibit some small but statistically significant differences. In Section 4, we show that our survey-based results are robust to controlling for these small imbalances.

Administrative data. We linked the student survey data to a rich set of individual-level administrative data covering the universe of high school students enrolled in the high schools of the Paris region over the period 2012/13 to 2016/17. These data provide detailed information on students' socio-demographic characteristics and enrollment status every year, allowing us to identify the high school track taken by Grade 10 students entering Grade 11.

The college enrollment outcomes of students in Grade 12 were obtained by matching the survey and administrative data for high school students with administrative microdata covering almost all students enrolled in selective and non-selective higher education programs in 2016/17.[16] These data are complemented with comprehensive individual examination results from the *Diplôme National du Brevet* (DNB), which is taken at the end of middle school, and from the national *Baccalauréat* exam (for Grade 12 students). Specifically, we use students' grades on the final exams in French and math (converted into national percentile ranks), as these tests are graded externally and anonymously. Further details about the data sources and the classification of higher education programs can be found in Appendix E.

was set randomly.

[16]Programs not covered by these administrative data are those leading to paramedical and social care qualifications. Available estimates suggest that among Grade 12 students who obtained a *Baccalauréat* in Science in 2008, under 6 percent were enrolled in such programs the following year (Lemaire, 2018).

3.2 Empirical Strategy

Compliance with random assignment was not perfect: about 5 percent of the classes assigned to the treatment group were not visited by a role model while 1 percent of the classes in the control group were mistakenly visited (see Appendix Table F7).[17] To deal with this marginal two-way non-compliance, we follow the standard practice of using treatment assignment as an instrument for treatment receipt, which allows us to estimate the program's local average treatment effect (LATE) instead of the average treatment effect (ATE). Specifically, we estimate the following model using two-stage least squares (2SLS):

$$Y_{ics} = \alpha + \beta D_{cs} + \theta_s + \epsilon_{ics}, \tag{1}$$

$$D_{ics} = \gamma + \delta T_{cs} + \lambda_s + \eta_{ics}, \tag{2}$$

where Y_{ics} denotes the outcome of student i in class c and high school s, D_{cs} is a dummy variable indicating whether the student's class received a visit, and T_{cs} is a dummy for assignment to the treatment group. School fixed effects, θ_s and λ_s, are included to account for the fact that the randomization was stratified by school and grade level.

The model described by Equations (1) and (2) is estimated separately by grade level and gender, and standard errors are clustered at the unit of randomization level (class). The results for the individual components of synthetic indexes are accompanied by adjusted p-values (q-values) in addition to the standard p-values, to account for multiple hypothesis testing.[18]

4 Effects of Classroom Interventions

We analyze the impact of the program on three main sets of student outcomes: (i) general perceptions of science-related careers and of gender roles in science; (ii) preferences and self-concept; and (iii) enrollment outcomes and academic performance.

The first set of outcomes captures students' representations of science-related studies and careers, as well as their beliefs regarding the underrepresentation of women in science and its possible causes. The program targeted these perceptions directly, in that the interventions were

[17] We are confident that non-compliance was mostly due to organizational and logistical issues and was not an endogenous response to randomization. The role models who ended up carrying out interventions in classes assigned to the control group or in classes that had not been selected to participate in the evaluation generally reported that their interventions had been poorly organized at the school level, with the person in charge often not being aware of the purpose of the visit. In some cases, classroom interventions were scheduled during another specialty course involving multiple classes, meaning that only some of the students in the treatment group were effectively treated.

[18] We use the False Discovery Rate (FDR) control, which designates the expected proportion of all rejections that are type-I errors. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008).

designed to convey non-stereotyped information about science-related careers and the place of women in science. We then investigate whether the program affected students' self-perceptions, changing their attitudes to the science subjects in their curriculum, their self-confidence in math, and their science-related career aspirations. Finally, we analyze whether the role models' interventions induced behavioral responses on the part of students, by examining their effects on enrollment decisions and academic performance.

4.1 Perceptions of STEM Careers and Gender Roles in Science

Students' post-intervention survey responses show that the classroom interventions were effective in challenging stereotyped views of science-related careers and gender roles.

Perceptions of science-related careers. Students were asked to agree with or disagree with five statements on science-related careers relating to pay, the length of studies leading to these careers, work-life balance, and the two commonplaces that science-related jobs are monotonous and solitary. We build a composite index of "positive perceptions of science-related careers" by re-coding the Likert scales so that higher values correspond to less stereotypical or negative perceptions, before taking the average of each student's responses to the five questions. To facilitate interpretation, we normalize the index to have a mean of zero and a standard deviation of 1 in the control group. For closer investigation of the various aspects that might be captured by the overall index, we further construct binary variables taking value 1 if the student agrees strongly or somewhat with each statement, and zero if he/she disagrees strongly or somewhat.[19]

The results are reported in Table 3. Students' baseline perceptions indicate relatively widespread negative stereotypes about careers in science (see columns 1 and 4), with little difference between boys and girls. About 30 percent of Grade 10 students in the control group consider that jobs in science are monotonous or solitary and that they are hard to reconcile with a fulfilling family life. More than 80 percent say that these jobs necessarily require long years of study, and over a third disagree with the statement that they pay more. Grade 12 students in the science track have slightly more positive perceptions, but the differences appear limited when balanced against the fact that these students have already self-selected into the science track.

As shown in Table 3, the role model interventions significantly improved girls' and boys'

[19]Similar groupings are performed when using responses that are measured on a four-point Likert scale (usually concerning perceptions or self-confidence) so that the outcome variables can be directly interpreted as proportions. We have checked that the results are not qualitatively affected by such grouping.

perceptions of science-related careers as measured by the composite index, in both Grade 10 and Grade 12. The effects range from 15 percent of a standard deviation for boys to around 30 percent for girls, with significantly greater effects for female students in both grades.[20] A significant impact of the program is observed for almost all the components of the index. The largest effects are found for the statements "science-related jobs require long years of study" and "science-related jobs are rather solitary," which embody two stereotypes that were specifically debunked in the slides and videos. Although the effects are not strikingly different between genders and grade levels, they do tend to be somewhat greater for girls in Grade 12. In particular, the program appears to have closed the gender gap in Grade 12 students' awareness of the earnings premium attached to science-related jobs, and to have reinforced girls' perception that such careers are compatible with a fulfilling family life.

Perceptions of gender roles in science. Female underrepresentation in STEM can be broadly attributed to three possible causes: gender differences in abilities, discrimination (on the demand side), and differences in preferences and career choices (on the supply side). The survey questions were designed to capture students' views on these dimensions.

Table 4 reveals the striking fact that more than a third of Grade 10 students and a quarter of Grade 12 students in the control group are not aware that women are underrepresented in science-related careers. These proportions are surprisingly similar by gender and by grade: one might well have expected girls in the science track in Grade 12 to be better informed of the gender imbalance in STEM. For boys and girls in both grades, we find that the interventions increased awareness of female underrepresentation in STEM by 12 to 17 percentage points. This is, perhaps predictably, one of the outcomes most strongly affected by the program.

The classroom interventions were also effective in debiasing students' beliefs about gender differences in math aptitude. To capture this dimension, we asked students whether they agreed with the statements that "men are more gifted than women in mathematics" and that "men and women are born with different brains." We used these two questions to construct a composite index to gauge whether students believe that men and women have equal aptitude for mathematics. The results show significant rises in this index for both genders in both grades, with treatment effects ranging between 9.5 percent and 14.8 percent of a standard deviation.[21]

Interestingly, the program had more ambiguous, partially unintended effects regarding the other two explanations. First, when asked about gender differences in preferences, the share

[20] Intention to treat (ITT) estimates, which are not reported in the main text, are very close to the LATE estimates due to almost perfect compliance with treatment assignment. They can be found for all main outcomes in Appendix Table I21.

[21] The detailed results for the two components of this index are reported in Appendix Table G11.

of students saying that women like science less than men is relatively low in the control group (16 percent of girls and 20 percent of boys in Grade 10; 7 percent of girls and 15 percent of boys in Grade 12), but it increases substantially due to the program for both genders, by 4 to 10 percentage points. Second, the baseline shares of boys and girls who declare that women are discriminated against in science-related jobs are much larger (around 60 percent); these too increase for both genders as a result of the program, by 7 to 15 percentage points.

How to interpret these contrasting effects on students' perceptions of gender roles in science? One of the key messages conveyed by the program materials is that women are severely underrepresented in science-related careers despite having the same aptitudes as men.[22] The program slides and videos documented the underrepresentation of women in STEM and rejected differences in aptitude as a cause, while remaining agnostic on other possible factors. Since the role models had a good deal of freedom and extra time for their interventions, they might have conveyed other messages or shared personal experiences regarding gender discrimination in science careers, which could explain why their interventions reinforced the belief that women are discriminated against. It is most unlikely, by contrast, that female scientists invited to act as role models would have pushed the idea that women like science less than men. A more plausible interpretation is that the program's unintended effect on students' perceptions of gender differences arose as an effort to rationalize why there are so few women in science-related careers, making students more likely to agree with the simplistic view that "women like science less than men" and to subscribe to the idea that women face discrimination in science careers.

4.2 Stated Preferences and Self-Concept

We now turn to the effects of the program on students' stated preferences and self-perception. Specifically, we investigate whether the interventions affected boys' and girls' taste for science subjects, their self-concept in math, and their science-related career aspirations. Table 5 reports the estimated treatment effect along these three dimensions, based on the composite indices constructed from the questionnaire answers. Treatment effects for each of the index questions are reported separately in Appendix G.

Taste for science subjects. The program had no sizeable impact on students' enjoyment of science subjects at school (reported on a 0 to 10 Likert scale), i.e., math, physics-chemistry, and earth and life sciences, or on their self-reported taste for science in general (see Appendix

[22]The second video shown in class ("Are we all Equal in Science?") was designed to convince students that women and men have similar brains and are equally capable of succeeding in math and science.

Table G12). That is, the effects are statistically insignificant for both genders in Grade 10 and Grade 12 when measured using the composite index combining students' responses to the four questionnaire items (see Table 5). These findings are not particularly surprising, given that the interventions did not expose students to science-related content and were not specifically designed to promote interest in science.

Math self-concept. To measure the impact of the program on students' self-concept in mathematics, we use a composite index that combines students' responses to four questions: (i) their self-assessed performance in math; (ii) whether they feel lost when trying to solve a math problem; (iii) whether they often worry that they will struggle in math class; and (iv) whether they consider that they can do well in science subjects if they make enough effort.

Consistent with the literature, our sample exhibits large gender differences in self-concept in mathematics. In the control group, the value of the index is 43 percent of a standard deviation lower for girls than for boys in Grade 10, and 37 percent lower in Grade 12. Large gender differences are found for most of the items used in the construction of this index, in particular those related to math anxiety (see Appendix Table G13).

Despite being a light-touch intervention, the program did have some positive effect on students' self-concept in math (see Table 5). Although these effects are only found to be statistically significant for boys in Grade 12 when using the composite index, the program consistently reduced the probability of students reporting worry that they will struggle in math class.[23] Point estimates tend to be higher for boys than for girls in both grades, implying that the classroom interventions had no effect on the substantial gender gap in this area.

Science-related career aspirations. The choice of a science-related career path does not depend solely on students' taste for the science subjects taught at school. It also depends on their perceptions of the relevant jobs and the amenities they may provide, such as earnings, work/life balance, and the work environment, all of which were embodied by the role models. As the program is found to have improved students' perceptions of science-related careers significantly, one might expect students in the treatment group to be more likely to consider these careers for themselves, even if their enjoyment of science subjects was unaffected.

To measure the effects on students' aspirations for science-related careers, we use a composite index combining the responses to four questions: (i) whether the students find that some jobs

[23]For each group of students, the correction of p-values for testing across multiple outcomes (see Appendix Table G13) cannot rule out the possibility that the effects on math anxiety are due to chance alone. However, finding a significant effect for the same variable across all four groups of students, which is not accounted for by the multiple testing correction, is suggestive of a genuine effect.

in science are interesting; (ii) whether they could see themselves working in a science-related job later in life; (iii) whether they report being interested in at least one of six STEM jobs out of a list of ten STEM and non-STEM occupations;[24] and (iv) whether they consider salary level as an important factor in their career choice.

Female students in Grade 12 are the only group of students for which we find significant effects on these science-related career aspirations, the value of the composite index being 11 percent of a standard deviation higher in the treatment than in the control group (see the last row of Table 5). The more detailed results reported in Appendix Table G14 show that the interventions had significant positive effects on three of the four corresponding survey items for girls in Grade 12. In particular, girls in the treatment group are more likely to report that earnings are an important factor in their career choice, which is consistent with the program raising their awareness of the wage premium for STEM jobs.

4.3 Educational Choices and Academic Performance

Access to rich administrative data on students' educational outcomes enables us to extend the analysis beyond student perceptions to document the program's impact on educational outcomes as well.

High school track after Grade 10. Panel A of Table 6 shows that the program had no significant impact on Grade 10 students' choice of track in the academic year following the intervention, i.e., 2016/17. For both genders, the treatment effect estimates are close to zero, whether we consider enrollment in any STEM track or enrollment in the general and technological STEM tracks separately.[25] Consequently, the program did not alter the 21-percentage-point gender gap in the likelihood of pursuing STEM studies after Grade 10.[26]

These results are consistent with the previous finding that the interventions had no discernible impact on Grade 10 students' aspirations for science-related careers. More generally, several mechanisms can be put forward to interpret the lack of effects on the enrollment status of Grade 10 girls in the following year. First, the program did not seem well suited to increase

[24]The STEM occupations in the list were: chemist, computer scientist, engineer, industrial designer, renewable energy technician, and researcher in biology. The non-STEM occupations were lawyer, pharmacist, physician, and psychologist.

[25]The more detailed results presented in Appendix Table G15 show that the distribution of students across non-STEM tracks (Humanities and Social sciences) did not change significantly either.

[26]We find similar results when considering the study intentions that Grade 10 students self-reported in the post-treatment survey, suggesting that the lack of effects on enrollment outcomes is due to students' choices being unaffected rather than to schools being less likely to admit treated students in the science-oriented tracks. Results are available upon request.

the share of girls enrolling in the STEM technological tracks in Grade 10, where the female share is particularly low (17 percent, see Appendix Figure A1). As discussed below, the positive effects that we find on the STEM enrollment decisions of girls in Grade 12 are concentrated among the high achievers in math. In grade 10, such students are unlikely to be directed to the technological track, explaining the lack of effects along this margin. Turning to the general science track, female underrepresentation is only moderate in Grade 11 (47 percent of students are girls, see Appendix Figure A1) and this track is the most common (usually the default choice) for high-performing students, including girls. Indeed, unlike the other high school tracks, it gives access to almost all fields of study in higher education. Female students who turn away from the science track in high school are therefore unlikely to consider a STEM career as a viable option, making their choices less easily reversible.[27]

Field of study after Grade 12. A central finding of the study is that the role model interventions had significant effects on the educational choices of girls in Grade 12, but not on those of their male classmates.

Panel B of Table 6 shows that for girls in Grade 12, the program increased the probability of enrolling in a STEM undergraduate program in 2016/17 by 2.4 percentage points (significant at the 10 percent level), which corresponds to an 8.3 percent increase from the baseline of 28.9 percent. The effect for boys is negligible and not statistically significant, implying that the gender gap in STEM enrollment narrowed from a baseline of 18.1 to 16.0 percentage points, i.e., an 11.6 percent reduction.[28]

As emphasized in Section 1.2, female underrepresenation in selective and male-dominated STEM fields account for approximately half of the STEM-related gender pay gap in France. Importantly, our results show that the program's positive impact on STEM enrollment is driven by a significantly larger fraction of girls in Grade 12 enrolling in both types of programs. The classroom interventions led to a highly significant 3.5 percentage-point increase in the fraction of girls enrolling in selective STEM programs, which represents a 32 percent increase from the baseline of 11.0 percent. The corresponding estimates for boys suggest that the classroom visits may have slightly increased male enrollment in these programs as well (by 2.0 percentage points from a baseline of 23.2 percent), but the effect is not statistically significant. Moreover, we

[27]Consistent with this interpretation, the survey data indicate that among Grade 10 students in the control group, only 22 percent of girls who did not enroll in the Grade 11 science track the following year declare that they could see themselves working in a science-related job, compared to 87 percent among those who did. In Grade 12, the gap is much less pronounced: the proportions are 64 percent among girls who did not enroll in a STEM undergraduate program the following year and 90 percent among those who did.

[28]With the caveat that we lack the statistical power to detect a significant reduction in the gender gap in STEM enrollment.

show in Section 4.4 that the magnitude of this effect for boys is substantially lessened when we control for students' baseline characteristics, suggesting that it probably depends on small residual imbalances in the male sample.[29]

Turning to the effects on enrollment in male-dominated STEM programs (mathematics, physics, computer science, and engineering), we find that the proportion of girls enrolling increased by a statistically significant 3.8 percentage points from a baseline of 16.6 percent (i.e., a 23 percent increase), compared to a non-significant 1.7-point increase for boys from a baseline of 37.9 percent. These results are particularly striking given that selective and male-dominated STEM programs are not only the most prestigious tracks but also those where the gender gap in enrollment is greatest. A simple back-of-the-envelope computation suggests that if our estimates could be extrapolated to the population of science-track Grade 12 students without considering general equilibrium effects, the female share would increase from 30 to 32 percent in STEM programs altogether, from 30 to 34 percent in selective STEM programs, and from 26 to 29 percent in male-dominated STEM programs.

Our estimates indicate that, on average, the role model interventions induced one girl in every two Grade 12 science-track classes to switch to a selective or a male-dominated STEM program at entry into higher education.[30] The more detailed results presented in Appendix Table G16 indicate that these effects are driven by female students shifting from non-STEM and female-dominated STEM programs. A significant decline in female enrollment is found for non-selective undergraduate programs in earth and life sciences (-2.2 percentage points), while small and non-significant reductions in the range of 0.4 to 0.8 point are found for selective programs in humanities and for non-selective programs in medicine, law and economics, humanities and psychology, and sports.

Taken together, the results for Grade 10 and Grade 12 students show that the program was only effective in steering girls towards the STEM tracks in which they are heavily underrepresented, even though two-thirds of the role models come from female-dominated STEM fields (earth and life sciences) and that the program was designed to promote all types of STEM careers, including those where women now outnumber men. These findings suggest that in the current setting, the role models affect only the most stereotyped choices.

[29]The balancing tests performed separately by grade level and gender do not point to unusually large imbalances between the treatment and control groups in any of the subsamples (see Appendix Tables F5 and F6). However, we do find that the predicted probability of being enrolled in a selective STEM program is marginally yet significantly higher (by 0.9 percentage point) in the treatment group than in the control group for boys in Grade 12.

[30]This computation is based on an average of 15 girls per class and an estimated 3.5 (respectively 3.8) percentage-point increase in the probability of enrolling in a selective (respectively male-dominated) STEM program.

Academic performance. The effects of the program on academic performance can be documented for students in Grade 12 based on the *Baccalauréat* exams, taken a few months after the classroom interventions (see Appendix Table G17). The treatment effect estimates on students' performance on the math test and on the probability of obtaining the *Baccalauréat* are close to zero and statistically insignificant for both genders. Although the role models could, in principle, have strengthened students' motivation to be admitted to the most selective STEM programs, resulting in their dedicating more time to studying mathematics and other science subjects, we find no evidence of any such effect. We can therefore rule out that the program's impact on the enrollment outcomes of girls in Grade 12 was driven by increased effort and accordingly better academic performance.

4.4 Robustness Checks

We conducted a number of robustness checks for our main findings (see Appendices H and I).

First, we investigated whether our treatment effect estimates for the survey-based outcomes might not be contaminated by the small imbalances in the response rates and observable characteristics of the treatment and control groups (see Section 3). We show that the estimated effects on students' perceptions are barely affected when controlling for students' observable characteristics (Appendix Table H18) and when weighting observations by the inverse of their predicted probability of answering the survey given their observable characteristics (Appendix Table H19).

Second, controlling for students' observable characteristics hardly affects the estimated effects on enrollment outcomes (see Appendix Table H20). If anything, the small positive (but not significant) effect on selective STEM enrollment for boys in Grade 12 becomes negligible.

Third, we checked whether our results are robust to using non-parametric randomization rather than model-based cluster-robust inference tests. The tests are performed by comparing our ITT estimates with the distribution of "placebo" ITT estimates obtained by randomly re-assigning treatment 2,000 times among participating classes within each school and grade level. The results yield empirical p-values that are generally close to the model-based p-values (see Appendix Table I21). Although they tend to be slightly more conservative, they confirm the program's statistically significant effects on female enrollment in selective and male-dominated STEM programs among Grade 12 students.

5 Persistence, Timing of Visits, Spillovers

This section extends the analysis to the persistence of effects on student perceptions, the timing of the interventions, and potential spillover effects on enrollment outcomes.

Persistence. The effects of the program on students' perceptions could be short-lived. We explore this issue by comparing the magnitude of treatment effects for different intervals between the intervention and the post-treatment survey: 1-2 months, 3-4 months and 5-6 months (see Appendix Table J22). The limited sample for each interval—especially those after 5-6 months—and the possibility that the quality of the interventions may have changed over time are two limitations that call for caution in drawing firm conclusions about the persistence of effects. With these caveats in mind, the results suggest that the treatment effects did not vanish quickly, insofar as they remain statistically significant for most outcomes beyond the first two months. The effects were, therefore, sufficiently persistent to affect students' choice of study.

Timing of visits. Earlier interventions seem to have had greater effects on the college choices of Grade 12 students, which could be made up to the end of May (see Appendix Figure J6). We find that classroom visits that took place in November increased female enrollment in selective or male-dominated STEM programs by 7 to 9 percentage points, compared with 3 to 6 points for visits in December-January and non-significant effects for visits in February-March.[31] These findings provide suggestive evidence that interventions made when many students are still undecided about their field of study and career plans may be more effective than those on the eve of irreversible choices.

Spillovers. An important issue is whether the program could have influenced the educational choices of students in the control group. These students may have heard about the visits directly, through their schoolmates in treatment group classes, or indirectly, through regular social interactions. If the direction of such effects is the same for students in the treatment and control groups, ignoring spillovers would cause us to underestimate the treatment effects.

The survey evidence suggests that the scope for spillover effects was limited, which is consistent with the notion that in the French school system most peer interactions take place within the class (Avvisati et al., 2014). In the treatment group, 58 percent of Grade 10 students

[31] The difference between the effects of visits before and after February 1 is statistically significant at the 5 percent level for girls and is robust to controlling for possible improvement or decline in the quality of role models' interventions over time, through the inclusion of fixed effects for the chronological order of the role models' classroom visits, i.e., first, second, etc.

and 63 percent of Grade 12 students report having talked about the classroom intervention with their classmates, but they are only 24 percent and 27 percent to report having talked with schoolmates from other classes, respectively (see Appendix Table K23).[32] In the control group, only 14 percent of students in Grade 10 report having heard of the classroom visits, mostly in a vague manner (12 percent). In Grade 12, students in the control group are more likely (34 percent) to report being at least vaguely aware of the visits, but under 5 percent of boys and girls have a precise recollection. Overall, these summary statistics suggest that spillover effects were quite limited.

We complement this survey evidence by investigating more formally whether the interventions affected the higher education choices of Grade 12 students whose classes were not assigned to the treatment group—either classes not selected by principals for the program or participating classes randomly assigned to the control group. Our empirical strategy, described in detail in Appendix K, builds on the following intuition: for schools that participated in the evaluation, the random assignment of treatment to participating classes makes it possible to estimate the average outcome that would have resulted if *all* students had only been exposed to the spillover effects of classroom interventions without being *directly* exposed to a role model. This unobserved "spillover-only" counterfactual can be estimated at the school level by computing an appropriately weighted average of the outcome of students in the non-participating classes and in the participating classes that were assigned to the control group.[33] The spillover effects of role model interventions are then estimated by comparing the "spillover-only" counterfactual to a "no-treatment" counterfactual. This second counterfactual is constructed using non-participating schools, which we observe in the administrative data, that have similar observable characteristics as the participating ones over the period 2012–2015. Having verified that trends in student enrollment outcomes were parallel between the two groups of schools in the pre-treatment period, we implement a difference-in-differences estimator to identify the program's spillover effects on students' STEM enrollment outcomes at college entry.

The results based on this difference-in-differences approach show no evidence of significant spillover effects of classroom visits on non-treated Grade 12 students (see Table K24 in the Appendix). Together with the survey evidence, they suggest that spillovers between treatment and control classes were at most limited.

[32]Interestingly, these proportions are higher for girls than for boys in the treatment group: 66 (70) percent of girls in Grade 10 (Grade 12) report having discussed the program with their classmates, 28 (33) percent with schoolmates from other classes, and 25 (27) percent with other students outside the school, compared with 50 (56) percent, 20 (21) percent, and 16 (13) percent for boys, respectively.

[33]This second group is given a greater weight to make it representative of both treated and non-treated participating classes, which, by virtue of randomization, have similar characteristics.

6 How Do Role Models Affect Student Behavior?

This section inquires into why light-touch classroom interventions by female role models with a background in science can affect girls' choice of study at university. Our insights are derived from comparison of groups of students who were exposed to different role models or who responded differently to a given role model. Two main conclusions emerge. First, there is considerable heterogeneity in the role models' ability to steer girls toward STEM studies, which highlights the importance of the individual who bears the message. Second, the role models who had the strongest impact on girls' choices were more effective in projecting a positive image of science-related careers and in stimulating students' aspirations for them, while putting less emphasis on the underrepresentation of women in science. These results suggest that some levers are more important than others in triggering a behavioral response. This offers useful indications for designing effective role model interventions.

The inquiry proceeds in four steps. First, we show that the treatment effects on STEM enrollment outcomes vary widely according to the two most obvious dimensions of heterogeneity in the current setting, namely students' academic performance and role models' background (professionals employed by L'Oréal vs. young researchers).

Using the approach developed by Chernozhukov et al. (2018), we then provide a more systematic analysis of the heterogeneity of treatment effects using machine learning techniques. The results confirm that the two dimensions hypothesized are also detected by an agnostic algorithm alongside other sources of heterogeneity.

Third, we shift attention from students' choice of study (the *final* or *behavioral* outcome) to their perceptions, self-concept, and interest for science (the possible *channels of influence*). Machine learning methods are also applied to detect the heterogeneity in treatment effects on these potential mediators. We identify the characteristics of the students and role models for whom we observe particularly large (or small) treatment effects both on the final outcome and on the possible channels. While this does not allow causal claims, it does offer useful insights into plausible channels through which role models affect female students' choice of study.

In the fourth step, we propose a method to generalize the results obtained in the third step so as not to depend on comparison of specific observable characteristics. We build on the approach proposed by Chernozhukov et al. (2018) to estimate the correlations between individual-level treatment effects on different outcomes conditional on exogenous observable characteristics. We seek to determine whether the students who were particularly receptive or unreceptive to some of the program's messages are also those whose choice of study was most

or least affected by the interventions. We find that the treatment effect on female enrollment in selective STEM programs is greater when the interventions resulted in larger improvements in girls' perceptions of science-related careers, and smaller when they raised the awareness of the underrepresentation of women in science. To the best of our knowledge, this approach constitutes an original methodological contribution that can serve to investigate channels of influence in other contexts.

6.1 Heterogeneous Treatment Effects on STEM Enrollment

We start by investigating how the treatment effects on STEM enrollment vary with the two most obvious dimensions of heterogeneity, i.e., math performance and role model background. Our analysis focuses on Grade 12 students, as we find no evidence of significant effects on STEM enrollment for Grade 10 students.[34]

High vs. low achievers in math. Applicants' performance in mathematics is the single most important admission criterion of selective undergraduate STEM programs. Using Grade 12 students' national percentile rank on the the *Baccalauréat* math test to proxy for academic performance, we find that the program's positive impact on selective STEM enrollment is driven by female students in the top quintile (see Figure 3 and Panel A of Appendix Table L27).[35] For these students, the probability of enrolling in a selective STEM program after high school increases by 16.3 percentage points, which corresponds to a 57 percent increase from the baseline of 28.5 percent. While the program also appears to have induced some male students in the top quintile to enroll in selective STEM programs, the effect is smaller (9.6 percentage points, or a 20 percent increase over the baseline of 49.2 percent) and only marginally significant. Especially striking is the fact that among the 20 percent top achievers in math, the gender gap in the probability of enrolling in a selective STEM program is the largest (21 percentage points) and the treatment reduces it by 6.7 percentage points, which corresponds to a 32 percent reduction from the baseline.[36]

[34]For the sake of completeness, the results of the heterogeneity analysis by level of performance in math and role model background for Grade 10 students can be found in Panel A of Appendix Tables L25 and L26.

[35]As discussed in Section 4.3, we find no significant impact of the program on students' performance on the math test of the *Baccalauréat* exam, which mitigates concerns about potential endogenous selection bias when conditioning on this variable. An alternative is to proxy Grade 12 students' math performance using their score on the DNB exam, taken at the end of Grade 9. In this case the results are qualitatively similar but less precise, presumably because DNB scores are a relatively noisy predictor of math performance in Grade 12.

[36]The differences in treatment effects between high and low achievers in math are qualitatively similar for enrollment in male-dominated STEM programs or all types of STEM programs (see Panel A of Appendix Table L27).

Role model background: researchers vs. professionals. It is unclear, a priori, how the different types of role models differ in their effect on students' attitudes and behavior. As shown in Table 2, role models with a research background are, on average, younger than the professionals employed by the sponsoring firm, which may foster a stronger sense of identification by the students. But because they work in highly specialized fields and in very competitive environments, it is not clear how attainable students might think their achievements are. On the other hand, the professionals tend to have higher pay and more experience, and come less often from a purely academic background. They also hold permanent positions, unlike Ph.D. candidates and postdocs.

Even though the role models were not randomly assigned to schools, we find no evidence of systematic differences in the observable characteristics of the students exposed to the two types (see Appendix Table F10), so we are confident that the heterogeneous treatment effects according to role model background are not confounded by differences in the characteristics of the classes they visited.

We find clear evidence that the two groups of role models had contrasting effects on STEM enrollment outcomes for girls in Grade 12. The left panel of Figure 4 shows that the professionals increased female students' probability of enrolling in a selective STEM program by a significant 5.3 percentage points, whereas researchers had no detectable effect.[37] The contrast is qualitatively similar, although less pronounced, when we consider enrollment in male-dominated STEM programs (right panel of Figure 4) or across all STEM programs (Appendix Table L27, Panel B). While the estimates also point to larger effects for boys who were exposed to role models with a professional background, they are not statistically significant at conventional levels.

6.2 Machine Learning to Uncover Sources of Heterogeneity

Investigating treatment effect heterogeneity by splitting the sample into subgroups inevitably entails the risk of data mining. To address this concern, we carry out a systematic exploration of treatment effect heterogeneity using machine learning (ML) methods (see Athey and Imbens, 2017, for a review). Specifically, we adopt the approach recently developed by Chernozhukov et al. (2018), as it appears well-suited to our objectives. A brief description is given below; a more detailed discussion can be found in Appendix M.[38]

[37] The difference between the treatment effects of the two groups of role models on Grade 12 girls' probability of enrolling in a selective STEM program is significant at the 5 percent level.

[38] So far, we are aware of only one other application of this method in ongoing work by Crépon, Duflo, Pariente, Seban, and Veillon. There are, however, several recent studies that use alternative machine learning methods to analyze treatment effect heterogeneity. See Bertrand et al. (2017) for a recent example based on the method developed by Wager and Athey (2018).

General description of Chernozhukov et al. (2018)'s approach. Let $Y(1)$ and $Y(0)$ denote the potential outcomes of a student when her class is and is not visited by a role model, respectively. Let Z be a vector of covariates that characterize the student and the role model who visited the class. The conditional average treatment effect (CATE), denoted by $s_0(Z)$, is defined as:

$$s_0(Z) \equiv \mathbb{E}[Y(1) - Y(0)|Z].$$

Because it is hard to obtain uniformly valid inference on the CATE without making strong assumptions, the approach in Chernozhukov et al. (2018) consists in conducting inference on specific *features* of the CATE, such as the expectation of $s_0(Z)$ in groups defined using a given ML predictor $S(Z)$. The first feature examined is the Best Linear Predictor (BLP) of $s_0(Z)$ given $S(Z)$. The authors show that the BLP can be identified from the following weighted linear projection:

$$Y = \alpha_0 + \alpha B(Z) + \beta_1(T - p(Z)) + \beta_2(T - p(Z))(S(Z) - \mathbb{E}[S(Z)]) + \epsilon, \quad E[w(Z)\epsilon X], \qquad (3)$$

where T is a dummy for treatment assignment; $B(Z)$ is an ML predictor of $Y(0)$ obtained from the training sample; $p(Z)$ is the probability of being treated conditional on the covariates Z;[39] $w(Z) = \{p(Z)(1 - p(Z))\}^{-1}$ is the weight; and X denotes the vector of all regressors ($X \equiv [1, B(Z), T - p(Z), (T - p(Z))(S(Z) - \mathbb{E}[S(Z)])]$). This projection identifies the parameters $\beta_1 = \mathrm{E}[s_0(Z)]$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$, which can both be estimated using the empirical analog of Equation (3) (see Appendix M for details). We refer to β_1 and β_2 in the tables as the average treatment effect (ATE) and heterogeneity loading (HET) parameters, respectively. The key parameter of interest, β_2, is informative about the correlation between the true and the predicted CATE. It is equal to 1 if the prediction is perfect and to 0 if there is no treatment effect heterogeneity or if $S(Z)$ has no predictive power.

The main purpose of estimating β_2 is to check if the trained ML methods are able to detect heterogeneity in the treatment effect. If so, the ML predictor of the CATE can be used to identify groups of individuals with the smallest and largest treatment effects. Heterogeneity groups are constructed by sorting students in the estimation sample based on the value of $S(Z_i)$, the predicted value of each student's treatment effect given his/her observable characteristics Z_i. We consider the bottom and top quintiles of $S(Z_i)$ and report ITT estimates for both groups of students—a feature of the CATE called Sorted Group Average Treatment Effects (GATEs) in Chernozhukov et al. (2018). We then compare the distribution of observable characteristics in

[39] In our setting, this probability is one half for most Grade 12 students.

the two groups—a feature called Classification Analysis (CLAN).

Inference. To avoid overfitting, we estimate the features of the CATE given an ML predictor $S(Z)$ on an *estimation sample* that is distinct from the *training sample* used to obtain $S(Z)$. We follow Chernozhukov et al. (2018) in iterating this data-splitting process and reporting the medians of estimates and p-values over several splits. The nominal levels of p-values are further adjusted to guarantee uniform validity, which leads to fairly conservative inference.[40]

Practical Implementation. We consider five alternative ML methods to estimate the predictor $S(Z)$: Elastic Net, Random Forest, Boosted Trees, Neural Network with feature extraction, and a simple linear model.[41] To train these methods, we use as covariates Z three indicators for the education districts of Paris, Créteil, and Versailles, four indicators for students' socioeconomic background (high, medium-high, medium-low, and low), their age, their overall percentile rank in the *Baccalauréat* exam, their percentile ranks in the French and math tests of the exam, and a vector of 56 role model fixed effects.[42] We limit ourselves to only a few exogenous student characteristics because our main objective is to document treatment effect heterogeneity across the 56 role models. For each outcome, the best ML method for either the BLP or the GATEs targeting of the CATE is selected using the performance measures proposed by Chernozhukov et al. (2018).[43]

Results. Using this procedure on the sample of girls in Grade 12, we find that the Elastic Net outperforms the other ML methods in predicting heterogeneous treatment effects on selective STEM enrollment, while for enrollment in male-dominated STEM, a linear model performs best (see Appendix Table M31).[44]

We use the corresponding ML predictors to estimate the parameters of the best linear

[40]While we report these adjusted p-values in all tables, we show in Appendix M that they are conservative in the sense that they are upper bounds for the true (unknown) p-values. We also provide a theoretical example in which these upper bounds are reached, implying that they cannot be improved upon. We argue, however, that this theoretical example is unlikely to be met in practical applications.

[41]These methods are implemented in R using the `caret` package written by Kuhn (2008), while the general approach of Chernozhukov et al. (2018) is implemented by adapting the codes made available online by the authors (`https://github.com/demirermert/MLInference`, accessed on May 4, 2018).

[42]Each student in the control group is assigned to the role model who visited his or her school, so the role model fixed effects are defined for students in both the treatment and control groups.

[43]The best ML method for the BLP targeting of the CATE is the one that maximizes the correlation between the ML predictor $S(Z)$ and the CATE $s_0(Z)$ in the estimation sample. The best method for the GATEs targeting of the CATE is the one that maximizes the sum of squares of the estimated GATEs across the heterogeneity groups. See Appendix M for more details on these performance measures.

[44]Since our main objective is to better understand the mechanisms underlying the significant effects on STEM enrollment outcomes for girls in Grade 12, we focus our main text analysis on this group of students. For the sake of completeness, the machine learning results for boys in Grade 12 are reported in Appendix Table M33.

predictor of the CATE from the weighted linear projection described in Equation (3). The results are shown in Panel A of Table 7. The estimated ATEs of the program on Grade 12 girls' enrollment in selective or male-dominated STEM are very close to those reported in Table 6 by virtue of the randomization of the sample splits. Turning to heterogeneity, the coefficients on the HET parameter indicate that the ML predictors are strongly and significantly correlated with the CATE on enrollment in selective STEM but not in male-dominated STEM.

Estimates of the sorted group average treatment effects (GATEs) for the top and bottom quintiles of the predicted treatment effects $S(Z)$ are reported in Panel B. They confirm the considerable heterogeneity of treatment effects on selective STEM enrollment among Grade 12 girls, GATEs ranging from a small negative effect in the bottom quintile to a large and significant 13.9 percentage point effect in the top quintile. The lesser heterogeneity in the effects on enrollment in male-dominated STEM is also confirmed, with no statistically significant difference between the top and bottom quintiles of treatment effects.

Panel C describes the characteristics of the 20 percent most and least affected students. The main takeaway is that the ML agnostic approach strongly confirms that the treatment effects on selective STEM enrollment are greater for high-achieving girls in math and for those who were exposed to a professional rather than a researcher role model. Between the 20 percent most and least affected female students, the average gap in math performance rank is as much as 63 percentile ranks; the difference in the probability that the class was visited by a professional is 14.8 percentage points. The results are qualitatively similar for enrollment in male-dominated STEM, but the differences between groups are smaller, which is consistent with the previous finding of less heterogeneous treatment effects for this outcome.

The results in Panel C disclose heterogeneous effects along other dimensions. The 20 percent of girls with the largest treatment effects on selective STEM enrollment perform significantly better in French and are from higher socioeconomic backgrounds, compared with the least affected 20 percent. They are also less likely to have been exposed to role models who have children or who graduated in a male-dominated STEM field, and more likely to have been exposed to role models who participated in the program the year before. However, the fact that these characteristics are correlated both with students' math performance and with the role model being either a professional or a researcher makes it difficult to determine their specific contribution to treatment effect heterogeneity. As suggestive evidence, we performed a "horse race" by regressing enrollment in selective or male-dominated STEM on the interactions between the treatment group indicator and each of the characteristics listed in Panel C. The results, which are shown in Appendix Table L28, are consistent with the conclusion that math

performance and role models' professional background are the two main observable dimensions of heterogeneity in the treatment effects on selective STEM enrollment.[45]

6.3 Heterogeneous Treatment Effects on Potential Mediators

To help identify the mechanisms behind the heterogeneity of effects on selective STEM enrollment among Grade 12 girls, we start by comparing the characteristics of those with the largest and smallest treatment effects for each of the potential channels of influence studied in Section 4, namely general perceptions of science-related careers and gender roles in science, taste for science subjects, math self-concept, and science-related career aspirations.[46]

The results are reported in Table 8. For each potential channel, we compare the characteristics of students in the top and bottom quintiles of predicted treatment effects. We focus on the two main sources of heterogeneity in the effects on enrollment in selective STEM, i.e., student performance in math and exposure to a role model with a professional background.[47] For the sake of completeness, Appendix Table L30 gives the results obtained via a more traditional heterogeneity analysis, i.e., comparing the LATEs for different subgroups of female students based on math performance and on the background of the role model.[48] The conclusions are broadly consistent with those deriving from the ML procedure.

The first key finding is that professionals and researchers were equally effective in debunking stereotypes on gender differences in math aptitude, while they reinforced students' perceptions that "women like science less than men" and that "women face discrimination in science-related jobs" to a comparable extent. These results suggest that the "gender debiasing" component of the classroom interventions, which emphasized men's and women's equal predisposition for science, cannot explain, alone, why the program increased girls' enrollment in selective STEM; otherwise the two groups of role models would be expected to have similar effects for this outcome, which is not what we find.

[45]The results in Appendix Table L28 indicate that the effect of the program on selective STEM enrollment remains significantly greater for high-achieving girls in math when we interact the treatment group indicator with other student and role model characteristics. By contrast, differences in treatment effects by academic performance in French or by socioeconomic background are no longer significant for girls. Moreover, consistent with the finding that the observable characteristics of students are reasonably balanced between the schools visited by the professionals and the researchers (see Appendix Table F10), the regression results confirm that the stronger effects of professionals on selective STEM enrollment are robust to fully interacting the treatment group indicator with the characteristics of students and role models.

[46]Each outcome is summarized using the relevant composite index, except for students' perceptions of gender roles in science, which are measured along multiple dimensions.

[47]For each of the outcomes listed in Table 8, Appendix Table M31 reports the performance of the different ML methods in predicting treatment effect heterogeneity based on the BLP (Panel A) or the GATEs targeting of the CATE (Panel B). The heterogeneity loading parameter of the BLP and the GATEs associated with the best ML method are reported separately for each outcome in Appendix Table M32.

[48]The corresponding results for students in Grade 10 can be found in Appendix Table L29.

By contrast, Table 8 reveals that the professionals were better than the researchers at improving female students' perceptions of science-related jobs and stimulating their aspirations for such careers, while emphasizing less the underrepresentation of women. The heterogeneity analysis indicates that the girls whose perceptions of science-related careers improved the most due to the program had more often a professional as role model: compared to girls in the bottom quintile of treatment effects for this outcome, those in the top quintile are 19.2 percentage points more likely to have been visited by a professional, the difference being statistically significant at the 1 percent level. Professionals are similarly overrepresented among the role models who had the greatest effects on girls' taste for science subjects (22.7 percentage-point gap between the top and bottom quintile of treatment effects), and even more so among those who raised science-related career aspirations the most (38.9 percentage-point gap). The opposite holds for heterogeneous treatment effects on the importance of female underrepresentation in STEM: compared to the 20 percent of girls least affected for this outcome, the 20 percent most affected are 11.2 percentage points more likely to have been visited by a researcher.

Together, these results provide a first description of the role models who were the most effective in changing female students' stereotyped behaviors. In addition to conveying positive information on career paths, these role models succeeded in sparking genuine interest in science and science-related jobs without overemphasizing the consequences of gender stereotyping. These features are in line with the main mechanisms usually considered necessary for role models to work: generating a sense of fit while moderating the effects of stereotype threat.

The analysis of treatment effect heterogeneity by student math performance tends to confirm that the messages conveyed by professionals were more effective at influencing female students' choice of studies. Indeed, the students who were particularly receptive to these messages are also those for whom we find the strongest impact on STEM enrollment, i.e., high achievers in mathematics. Average math performance is significantly higher among the students whose perceptions of science-related careers and taste for science subjects improved the most. Conversely, we find fewer high achievers among the girls whose awareness of female underrepresentation in STEM and perception of gender discrimination increased the most.

While these comparisons on the basis of role model background and student math performance cannot be given a causal interpretation, they are consistent with the notion that gender-neutral messages about careers in science are more effective than gender-related messages to steer girls towards STEM studies. The next section provides additional evidence supporting this interpretation.

6.4 Correlation between Treatment Effects

So far, our discussion of the channels of influence has sought to identify the main dimensions of treatment effect heterogeneity on STEM enrollment outcomes and investigated how the impact on student perceptions varies along these dimensions. We will now develop a more general approach to estimate the correlation between the treatment effects on different outcomes. That is, given their observable characteristics, are the students with the largest treatment effects for a potential channel of influence Y^A the same ones who exhibit the largest treatment effects on enrollment outcome Y^B?

A new feature of the CATE. Because treatment effects for a given student are never observed, the correlation between *individual-level* treatment effects on outcomes Y^A and Y^B cannot be estimated without making strong assumptions.[49] Instead, our approach takes advantage of the predicted heterogeneity in treatment effects by student and role model characteristics to recover the correlation $\rho_{A,B|Z} = \text{Corr}(s_0^A(Z), s_0^B(Z))$ between the true CATEs on the two outcomes Y^A and Y^B, which we denote by $s_0^A(Z)$ and $s_0^B(Z)$, respectively. As discussed in Appendix M, we believe that this is an interesting alternative to other methods, such as causal mediation analysis, that are commonly used in the medical and social sciences literature to identify the factors that may be part of the causal pathway between an intervention and an outcome. Our proposed method does not depend on strong identifying assumptions and can be used in any experimental setting provided that there are enough observed exogenous covariates—a condition that is met in an increasing number of empirical studies.

To estimate the correlation between $s_0^A(Z)$ and $s_0^B(Z)$, we first define a new feature of the CATE as a simple adaptation of Chernozhukov et al. (2018)'s method. Instead of estimating the Best Linear Predictor of $s_0^A(Z)$ based on the ML predictor $S^A(Z)$, we estimate the BLP of $s_0^A(Z)$ based on $S^B(Z)$, i.e., the ML predictor of the heterogeneity in treatment effects on outcome Y^B. The heterogeneity loading parameter of the BLP we are interested in is

$$\beta_2^{A|B} = \text{Cov}(s_0^A(Z), S^B(Z))/\text{Var}(S^B(Z)). \tag{4}$$

This parameter is identified and can be estimated using a variant of Equation (3) (see details in Appendix M). By switching the roles of Y_A and Y_B in Equation (4), one can similarly estimate

[49] In Appendix M, we also argue that, due to sampling error, estimating separate treatment effects for each role model before computing the correlation between the estimated role-model-specific treatment effects for outcomes Y^A and Y^B in the same sample would likely result in a biased estimate of the true correlation between treatment effects.

the heterogeneity loading parameter from the BLP of $s_0^B(Z)$ based on $S^A(Z)$, i.e.,

$$\beta_2^{B|A} = \text{Cov}(s_0^B(Z), S^A(Z))/\text{Var}(S^A(Z)).$$

Writing $S^A(Z) = s_0^A(Z) + \eta_A$ and $S^B(Z) = s_0^B(Z) + \eta_B$ and assuming that the prediction errors η_A and η_B are independent of both predicted functions $s_0^A(Z)$ and $s_0^B(Z)$ in the estimation sample, we show that $\beta_2^{A|B}$ and $\beta_2^{B|A}$ have the same sign, which is indicative of whether the treatment effects on Y^A are positively or negatively correlated with the treatment effects on Y^B.

In Appendix M, we show that under these assumptions, the correlation between the true CATEs on Y^A and Y^B, $\rho_{A,B|Z}$, can be estimated using the following formula:[50]

$$\rho_{A,B|Z} = \text{Sign}(\beta_2^{A|B}) \frac{\sqrt{\beta_2^{A|B}\beta_2^{B|A}}}{\sqrt{\beta_2^{B|B}}\sqrt{\beta_2^{A|A}}}, \tag{5}$$

where $\beta_2^{A|A}$ and $\beta_2^{B|B}$ are the heterogeneity loading parameters in the BLPs of $s_0^A(Z)$ and $s_0^B(Z)$ on their respective predictors $S^A(Z)$ and $S^B(Z)$.

Practical implementation. As in the previous section, we split the data into a training and an estimation sample. We obtain predictors $S^A(Z)$ and $S^B(Z)$ of $s_0^A(Z)$ and $s_0^B(Z)$ in the training sample and use them to estimate the four parameters $\beta_2^{A|A}$, $\beta_2^{B|B}$, $\beta_2^{A|B}$ and $\beta_2^{B|A}$ in the estimation sample. We then plug these parameter estimates in Equation (5) to obtain an estimate $\hat{\rho}_{A,B|Z}$ of the correlation between the CATEs on outcomes Y^A and Y^B. We use a bootstrap procedure, also performed in the estimation sample, to obtain a 95 percent confidence interval for $\hat{\rho}_{A,B|Z}$.[51] As in the previous section, we follow the procedure of Chernozhukov et al. (2018) so that our final estimates of $\rho_{A,B|Z}$ and its confidence interval are computed as medians of estimates obtained from several estimation samples, the nominal level of confidence intervals being adjusted to guarantee uniform validity.[52]

[50] While it is not possible to prove that the out-of-sample prediction error of a ML predictor is independent from the predicted outcome for any predictor, this assumption seems reasonable when using efficient ML algorithms such as those considered in this paper. As suggestive evidence, we have checked in Monte Carlo simulations that this assumption holds for a large set of simulated functions of Z, which are generated manually and predicted on subsamples of our data. We further checked that the correlation $\rho_{A,B|Z}$ is successfully recovered for various data-generating processes using the formula in Equation (5).

[51] We report confidence intervals rather than p-values because the former are highly skewed, implying that the p-values obtained from bootstrap under normality assumptions are misleading.

[52] In theory, $\beta_2^{A|A}$ and $\beta_2^{B|B}$ should both be positive while $\beta_2^{A|B}$ and $\beta_2^{B|A}$ should have the sign of $\rho_{A,B|Z}$ in each iteration of the data-splitting process. However, this is not always the case in practice due to estimation error, in particular when the predictors $S^A(Z)$ and $S^B(Z)$ are very noisy. When one of the conditions above is not satisfied in a given estimation sample, we do not estimate $\rho_{A,B|Z}$ and discard the corresponding iteration of the data-splitting procedure. Reassuringly, the sensitivity analysis provided in in Appendix Table M35 shows that our results are barely affected when we exclude data splits that yield a poor ML prediction of the CATEs

Results. The results for girls in Grade 12 are reported in Table 9, where the covariates that we use to predict treatment effect heterogeneity are the same as in Table 7. They suggest that some channels were more important than others in steering female students towards STEM studies. The treatment effects on girls' enrollment in selective STEM exhibit a strong positive and significant correlation with the improvement in their perceptions of science-related careers ($\hat{\rho} = 0.96$) and with the improvement in their taste for science subjects ($\hat{\rho} = 0.71$). We also find evidence that girls who became more aware of the underrepresentation of women in science careers were less likely to change their choice towards a selective STEM program ($\hat{\rho} = -0.68$).

While not statistically significant at the 5 percent level, the remaining correlations give some indication on the role of other candidate channels.[53] They confirm in particular that debiasing girls' attitudes towards gender differences in aptitude for math is not associated with increased enrollment in selective STEM programs ($\hat{\rho} = 0.19$ with a 95 percent confidence interval of $[-1.24, 2.05]$) and that, if anything, reinforcing the belief that women are discriminated in science careers tends to deter girls from enrolling in selective STEM programs ($\hat{\rho} = -0.34$ $[-2.22, 0.56]$). By contrast, raising girls' aspirations for careers in science is associated with an increased probability that they enroll in such programs ($\hat{\rho} = 0.36$ $[-0.51, 2.01]$).

Overall, the results based on correlations between treatment effects are in line with and extend those obtained in the previous section. They suggest that the most effective role models were those who managed to convey a positive image of science careers without overemphasizing women's underrepresentation and its possible causes.

7 Conclusion and Discussion

Based on a large-scale randomized field experiment involving 56 female role models and nearly 20,000 high school students in Grade 10 and Grade 12, this paper shows that a one-hour in-class exposure to a female scientist can significantly increase female participation in STEM fields of study at college enrollment. Remarkably, the positive enrollment effects are observed only in the tracks with the most severe gender imbalance, which are the most prestigious and selective, and those that are most math-intensive. These effects can be expected to improve the future earnings of the target population, since the selective and male-dominated STEM programs offer high wage premia relative to other programs.

on outcomes Y^A or Y^B.

[53] We report in Table 9 the lower and upper bounds for the lower and upper limits of the actual 95 percent confidence interval associated with each estimated correlation. Recall that the (unknown) true confidence intervals are likely to be smaller than suggested by the bounds reported in Table 9 (see discussion in Appendix M).

In our empirical setting, the role model interventions had no discernable effects on students' taste for science or their academic performance, and only slightly improved their math self-concept, thus ruling out these factors as primary causes of the observed effects on STEM enrollment. By contrast, the classroom visits significantly challenged students' stereotyped views of science careers and gender differences in aptitude for science. These effects, however, are observed for both genders in both grades, suggesting that by themselves they cannot explain why the role model interventions only affected the educational choices of girls in Grade 12.

Our results offer substantial evidence that female students' behavioral response to the role model interventions was mediated by their ability to identify with the female scientists to whom they were exposed. On the verge of important decisions about their future education and career pathways, girls in Grade 12 appear to have been more receptive than the other groups of students to the attractive and de-masculinized image of science-related careers embodied by the role models. Consistently with this, we find that their improved perceptions of science careers translated into stronger aspirations for such careers. This process of identification was less likely to occur among Grade 10 girls, who are further away from career choices, and for boys in both grade levels, who may have found it more difficult to identify with women scientists. To confirm this latter hypothesis and, more generally, to improve our understanding of role model effects, an interesting avenue for future research would be to compare the impact of male and female role models in a similar context.[54]

Another important insight from the study is that by heightening awareness of the underrepresentation of women in STEM, while at the same time emphasizing men's and women's equal aptitude for science, the interventions may have unintentionally reinforced students' beliefs that women dislike science and face discrimination in STEM careers. That is, there is suggestive evidence that excessive stress on gender can be counter-productive and that gender-neutral messages might be more effective in steering girls towards STEM fields. In our setting, the role models who most reinforced the perception that women are underrepresented and discriminated against in science had the least effect on selective STEM enrollment for female students in Grade 12, whereas those who most improved girls' perceptions of science careers had the greatest impact. These conclusions echo those reached by Banerjee et al. (2013) in a very different context: using a large-scale randomized experiment to test whether citizens can learn from others' experiences about the quality of female leaders in India, the authors show that the vote share for the incumbent was more sensitive to past performance in places where a gender-neutral

[54]This aspect could not be investigated in this study, as the program was already under way, with only female role models.

campaign was run than where the "gender" theme was broached. These findings suggest that role model interventions need to be carefully designed to limit the potential discouragement effect of overemphasis on gender imbalances.

More generally, our heterogeneity analysis warns against the temptation to view role models as a one-size-fits-all remedy against female underrepresentation in STEM fields. Like Carrell et al. (2010), we find that role model effects on enrollment outcomes are concentrated among high-achieving girls in math. The effectiveness of this type of intervention in increasing female participation in STEM among lower-performing students remains an open question. Our study also highlights the importance of role models' profile in generating a sense of fit among students, as the effects on educational choices varied markedly across the participating female scientists. These results point to the need for further research on how the matching between role models and students can be optimized to make this particular type of intervention more effective.

References

Anderson, Michael L., "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*, 2008, *103* (484).

Ashraf, Nava, Natalie Bau, Corinne Low, and Kathleen McGinn, "Negotiating a Better Future: How Interpersonal Skills Facilitate Inter-generational Investment," *Quarterly Journal of Economics*, forthcoming.

Athey, Susan and Guido Imbens, "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 2016, *113* (27), 7353–7360.

_ **and Guido W. Imbens**, "The Econometrics of Randomized Experiments," in Esther Duflo and Abhijit V. Banerjee, eds., *Handbook of Economic Field Experiments*, Vol. 1, Elsevier, 2017, pp. 73–140.

Avvisati, Francesco, Marc Gurgand, Nina Guyon, and Éric Maurin, "Getting Parents Involved: A Field Experiment in Deprived Schools," *Review of Economic Studies*, 2014, *81* (1), 57–83.

Babcock, Linda and Sara Laschever, *Women don't Ask: Negotiation and the Gender Divide*, Princeton University Presss, 2012.

Banerjee, Abhijit V., Esther Duflo, Clément Imbert, and Rohini Pande, "Entry, Exit and Candidate Selection: Experimental Evidence from India," 2013. Manuscript, 3ie Grantee Final Report.

Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova, "Female Leadership Raises Aspirations and Educational Attainment for Girls: A policy Experiment in India," *Science*, 2012, *335* (6068), 582–586.

Beede, David, Tiffany Julian, David Langdon, George McKittrick, Beethika Khan, and Mark Doms, "Women in STEM: A Gender Gap to Innovation," 2011. U.S. Department of Commerce, Economics and Statistics Administration, Issue Brief No. 04-11.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli, "Adaptive Linear Step-up Procedures that Control the False Discovery Rate," *Biometrika*, 2006, *93* (3), 491–507.

Bertrand, Marianne, Bruno Crépon, Alicia Marguerie, and Patrick Premand, "Contemporaneous and Post-Program Impacts of a Public Works Program: Evidence from Côte d'Ivoire," 2017. Manuscript.

Bettinger, Eric P. and Bridget Terry Long, "Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students," *American Economic Review*, 2005, *95* (2), 152–157.

Betz, Diana E. and Denise Sekaquaptewa, "My Fair Physicist? Feminine Math and Science Role Models Demotivate Young Girls," *Social Psychological and Personality Science*, 2012, *3* (6), 738–746.

Black, Dan A., Amelia M. Haviland, Seth G. Sanders, and Lowell J. Taylor, "Gender Wage Disparities among the Highly Educated," *Journal of Human Resources*, 2008, *43* (3), 630–650.

Blau, Francine D. and Lawrence M. Kahn, "The Gender Wage Gap: Extent, Trends, and Explanations," *Journal of Economic Literature*, 2017, *55* (3), 789–865.

Breda, Thomas and Mélina Hillion, "Teaching Accreditation Exams Reveal Grading Biases Favor Women in Male-Dominated Disciplines in France," *Science*, 2016, *353* (6298), 474–478.

_ **and Son-Thierry Ly**, "Professors in Core Science Fields are not always Biased Against Women: Evidence from France," *American Economic Journal: Applied Economics*, 2015, *7* (4), 53–75.

Brown, Charles and Mary Corcoran, "Sex-Based Differences in School Content and the Male-Female Wage Gap," *Journal of Labor Economics*, 1997, *15* (3), 431–465.

Burgess, Simon, "Michelle Obama and an English School: The Power of Inspiration," 2016. Manuscript.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek, "Gender, Competitiveness, and Career Choices," *Quarterly Journal of Economics*, 2014, *129* (3), 1409–1447.

Canaan, Serena and Pierre Mouganie, "Female Science Advisors and the STEM Gender Gap," 2019. Manuscript.

Canes, Brandice J. and Harvey S. Rosen, "Following in her Footsteps? Faculty Gender Composition and Women's Choices of College Majors," *Industrial and Labor Relations Review*, 1995, *48* (3), 486–504.

Carlana, Michela, "Implicit Stereotypes: Evidence from Teachers' Gender Bias," *Quarterly Journal of Economics*, 2019, *134* (3), 1163–1224.

Carnevale, Anthony P., Nicole Smith, and Michelle Melton, *STEM: Science, Technology, Engineering, Mathematics*, Washington, DC: Georgetown University Center on Education and the Workforce, 2011.

Carrell, Scott E., Marianne E. Page, and James E. West, "Sex and Science: How Professor Gender Perpetuates the Gender Gap," *Quarterly Journal of Economics*, 2010, *125* (3), 1101–1144.

Ceci, Stephen J. and Wendy M. Williams, "Understanding Current Causes of Women's Underrepresentation in Science," *Proceedings of the National Academy of Sciences*, 2011, *108* (8), 3157–3162.

_ **, Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams**, "Women in Academic Science: A Changing Landscape," *Psychological Science in the Public Interest*, 2014, *15* (3), 75–141.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," 2018. NBER Working Paper No. 24678.

Cheryan, Sapna, John Oliver Siy, Marissa Vichayapai, Benjamin J. Drury, and Saenam Kim, "Do Female and Male Role Models who Embody STEM Stereotypes Hinder Women's Anticipated Success in STEM?," *Social Psychological and Personality Science*, 2011, *2* (6), 656–664.

Correll, Shelley J., "Gender and the Career Choice Process: The Role of Biased Self-Assessments," *American Journal of Sociology*, 2001, *106* (6), 1691–1730.

Dasgupta, Nilanjana 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*, 2004, *40* (5), 642–658.

Dee, Thomas S., "Teachers and the Gender Gaps in Student Achievement," *Journal of Human Resources*, 2007, *42* (3), 528–554.

Del Carpio, Lucia and Maria Guadalupe, "More Women in Tech? Evidence from a Field Experiment Addressing Social Identity," 2018. CEPR Discussion Paper DP13234.

Eble, Alex and Feng Hu, "Child Beliefs, Societal Beliefs, and Teacher-Student Identity Match," 2017. CDEP-CGEG Working Paper No. 43.

Ehrlinger, Joyce and David Dunning, "How Chronic Self-Views Influence (and Potentially Mislead) Estimates of Performance," *Journal of Personality and Social Psychology*, 2003, *84* (1), 5–17.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, 2003, *118* (3), 1049–1074.

Hoffmann, Florian and Philip Oreopoulos, "A Professor Like Me: The Influence of Instructor Gender on College Achievement," *Journal of Human Resources*, 2009, *44* (2), pp. 479–494.

Hoogendoorn, Sander, Hessel Oosterbeek, and Mirjam van Praag, "The Impact of Gender Diversity on the Performance of Business Teams: Evidence from a Field Experiment," *Management Science,*, 2013, *59* (7), 1514–1528.

Hyde, Janet S., "The Gender Similarities Hypothesis," *American Psychologist*, 2005, *60* (6), 581–591.

Kofoed, Michael S. and Elizabeth McGovney, "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*, 2019, *54* (2), 430–467.

Kuhn, Max, "Building Predictive Models in R using the caret Package," *Journal of Statistical Software*, 2008, *28* (5), 1–26.

Lavy, Victor and Edith Sand, "On the Origins of Gender Gaps in Human Capital: Short- and Long-Term Consequences of Teachers' Biases," *Journal of Public Economics*, 2018, *167* (C), 263–269.

Leibbrandt, Andreas and John A. List, "Do Women Avoid Salary Negotiations? Evidence from a Large-Scale Natural Field Experiment," *Management Science*, 2015, *61* (9), 2016–2024.

Lemaire, Sylvie, "Parcours dans l'enseignement supérieur: devenir des bacheliers 2008," 2018. Note d'Information 12.10, MESR-SIES.

Lim, Jaegeum and Jonathan Meer, "The Impact of Teacher-Student Gender Matches: Random Assignment Evidence from South Korea," *Journal of Human Resources*, 2017, *52* (4), 979–997.

_ **and** _ , "Stereotypes, Role Models, and the Formation of Beliefs," *Journal of Human Resources*, 2019, *54* (2), 430–467.

Lockwood, Penelope and Ziva Kunda, "Superstars and Me: Predicting the Impact of Role Models on the Self," *Journal of Personality and Social Psychology*, 1997, *73* (1), p. 91.

Marginson, Simon, Russell Tytler, Brigid Freeman, and Kelly Roberts, *STEM: Country comparisons*, Report for the Australian Council of Learned Academies, 2013.

Mullainathan, Sendhil and Jann Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Muralidharan, Karthik and Ketki Sheth, "Bridging Education Gender Gaps in Developing Countries: The Role of Female Teachers," *Journal of Human Resources*, 2016, *51* (2), 269–297.

National Science Foundation, *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2017*, National Science Foundation and National Center for Science and Engineering Statistics, 2017. Special Report NSF 17-310. Arlington, VA.

Neumark, David and Rosella Gardecki, "Women Helping Women? Role Model and Mentoring Effects on Female Ph.D. Students in Economics," *Journal of Human Resources*, 1998, *33* (1), 220–246.

Nguyen, Trang, "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar," 2008. Manuscript.

Niederle, Muriel and Lise Vesterlund, "Do Women Shy away from Competition? Do Men Compete too Much?," *Quarterly Journal of Economics*, 2007, *122* (3), 1067–1101.

_ **and** _ , "Explaining the Gender Gap in Math Test Scores: The Role of Competition," *Journal of Economic Perspectives*, 2010, *24* (2), 129–144.

O'Brien, Laurie T., Aline Hitti, Emily Shaffer, Amanda R. Van Camp, Donata Henry, and Patricia N. Gilbert, "Improving Girls' Sense of Fit in Science: Increasing the Impact of Role Models," *Social Psychological and Personality Science*, 2016, *8* (3), 301–309.

Paredes, Valentina, "A Teacher Like Me or a Student Like Me? Role Model Versus Teacher Bias Effect," *Economics of Education Review*, 2014, *39*, 38–49.

Porter, Catherine and Danila Serra, "Gender Differences in the Choice of Major: The Importance of Female Role Models," *American Economic Journal: Applied Economics*, forthcoming.

Riise, Julie, Barton Willage, and Willén Alexander, "Can Female Doctors Cure the Gender STEMM Gap? Evidence from Randomly Assigned General Practitioners," 2019. NHH DP No. 18/2019.

Rothstein, Donna S., "Do Female Faculty Influence Female Students' Educational and Labor Market Attainments?," *Industrial and Labor Relations Review*, 1995, *48* (3), 515–530.

Spelke, Elizabeth S., "Sex Differences in Intrinsic Aptitude for Mathematics and Science: A Critical Review," *American Psychologist*, 2005, *60* (9), 950–958.

Terrier, Camille, "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement," *Economics of Education Review*, forthcoming.

Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

Weinberger, Catherine J., "Mathematical College Majors and the Gender Gap in Wages," *Industrial Relations*, 1999, *38* (3), 407–413.

Xue, Yi and Richard C. Larson, "STEM Crisis or STEM Surplus? Yes and Yes," *Monthly Labor Review*, 2015, *138* (5), 1–13.

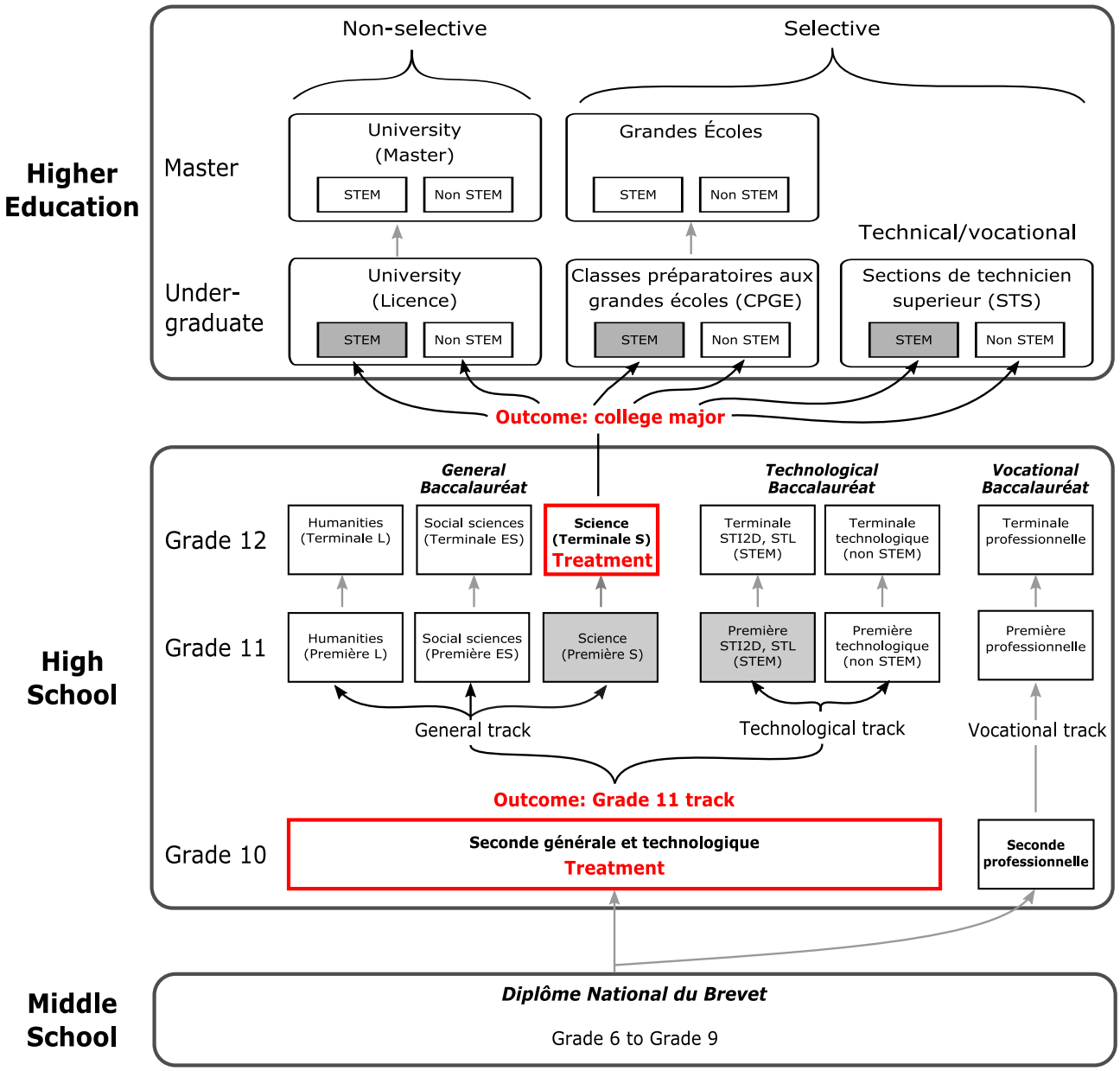


Figure 1 – Tracks in Secondary and Post-Secondary Education in France

Notes: The figure describes the structure of upper secondary and post-secondary education in France. The role model interventions took place in Grade 10 and Grade 12 (science track). Science-oriented high school tracks and STEM undergraduate programs are highlighted in grey.

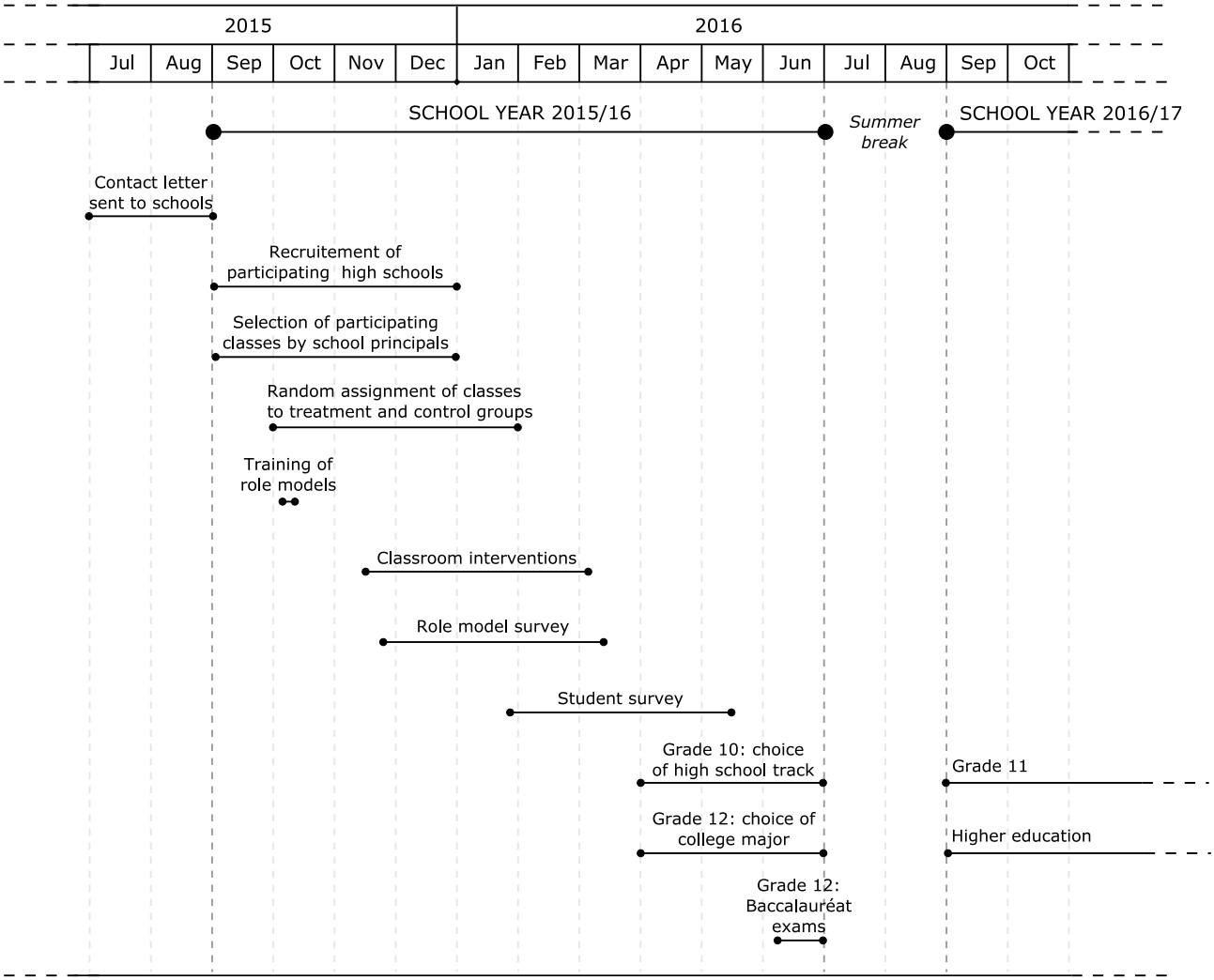


Figure 2 – Program Evaluation Timeline

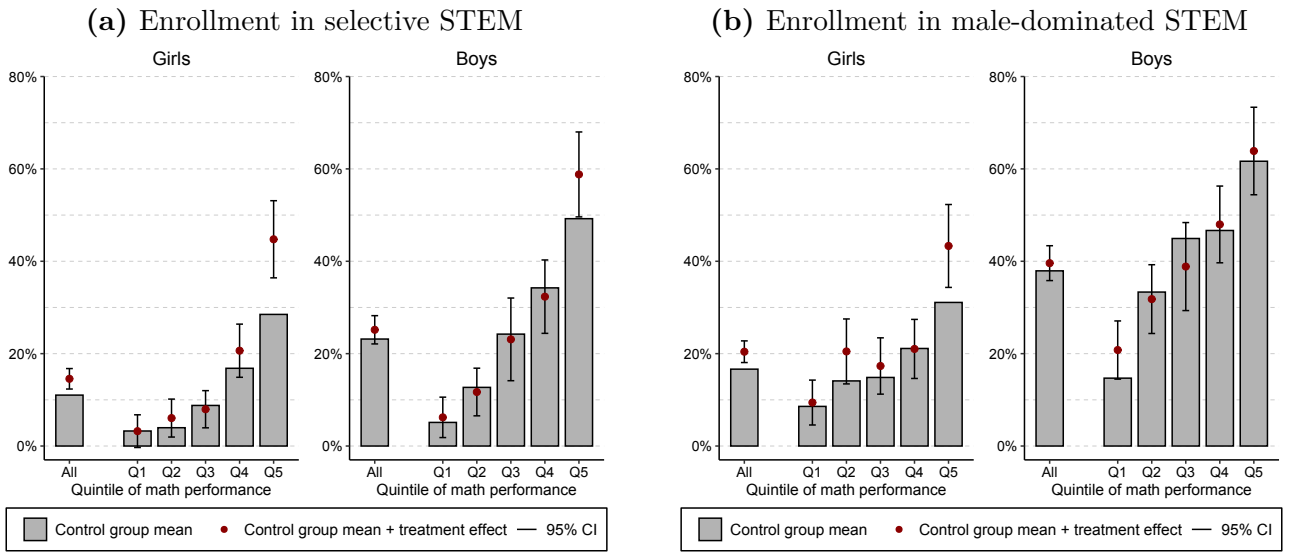


Figure 3 – Grade 12 Students: Enrollment in Selective and Male-Dominated STEM Undergraduate Programs, by Gender and Quintile of *Baccalauréat* Performance in Math

Notes: The figure shows the fraction of Grade 12 (science track) students enrolled in selective (Panel a) and in male-dominated (Panel b) STEM undergraduate programs in the year following high school graduation, separately for girls and boys. The filled bars indicate the baseline enrollment rates among students in the control group, both overall and separately by quintile of *Baccalauréat* performance in math. The solid circles show the estimated treatment effects (added to the control group means), with 95 percent confidence intervals denoted by vertical capped bars. The local average treatment effects are estimated from a regression of the outcome of interest on interactions between a classroom visit indicator and the quintiles of math performance, using treatment assignment (interacted with the quintiles of math performance) as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors are adjusted for clustering at the unit of randomization (class). The detailed results are provided in Appendix Table L27, Panel A.

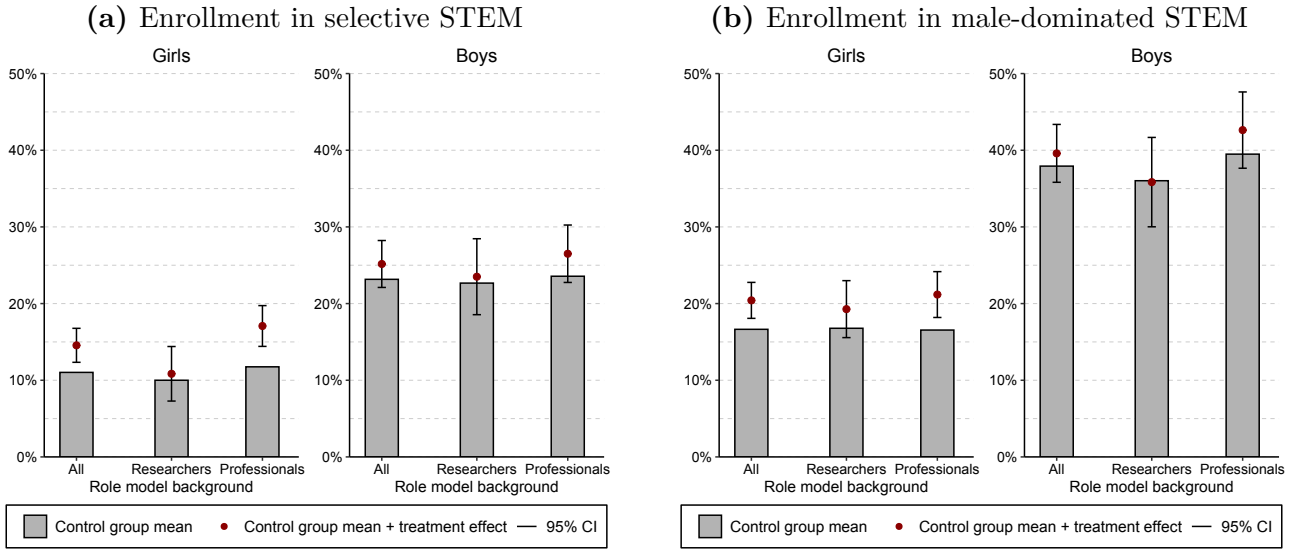


Figure 4 – Grade 12 Students: Enrollment in Selective and Male-Dominated STEM Undergraduate Programs, by Gender and Role Model Background

Notes: The figure shows the fraction of Grade 12 (science track) students enrolled in selective (Panel a) and in male-dominated (Panel b) STEM undergraduate programs after graduating from high school, separately for girls and boys. The filled bars indicate the baseline enrollment rates among students in the control group, both overall and separately by type of female role model who visited the classroom (researcher or professional). The solid dots show the estimated treatment effects (added to the control group means), with 95 percent confidence intervals denoted by vertical capped bars. The local average treatment effects are estimated from a regression of the outcome of interest on interactions between a classroom visit indicator and two indicators for role model type, using treatment assignment (interacted with role model type) as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors are adjusted for clustering at the unit of randomization (class). The detailed results are provided in Appendix Table L27, Panel B.

Table 1 – Treatment-Control Balance

	Control group (1)	Treatment group (2)	Within school: Difference T−C (3)	Within school: p-value of diff. (4)
Panel A. Grade 10				
Student characteristics				
Female	0.525	0.511	−0.010	0.309
Age (years)	15.73	15.71	−0.01	0.180
Non-French	0.057	0.057	0.002	0.652
High SES	0.399	0.412	0.008	0.321
Medium-high SES	0.133	0.128	−0.007	0.168
Medium-low SES	0.239	0.225	−0.012	0.064
Low SES	0.229	0.235	0.012	0.085
Number of siblings	1.492	1.486	0.003	0.904
Class size	32.79	32.95	0.07	0.476
At least one science elective course	0.529	0.540	0.005	0.820
At least one standard elective course	0.542	0.519	−0.031	0.138
DNB percentile rank in math	63.09	62.90	−0.35	0.533
DNB percentile rank in French	61.11	61.40	0.12	0.829
Test of joint significance	F-stat: 0.798 (p-value: 0.653)			
Predicted track in Grade 11				
Grade 11: Science track	0.369	0.370	−0.001	0.912
Grade 11: Science - general track	0.307	0.308	0.000	0.962
Grade 11: Science - technological track	0.061	0.061	0.000	0.865
N	6,801	6,899	13,700	
Panel B. Grade 12 (science track)				
Student characteristics				
Female	0.499	0.484	−0.014	0.292
Age (years)	17.14	17.11	−0.04	0.000
Non-French	0.053	0.048	−0.006	0.275
High SES	0.453	0.474	0.029	0.009
Medium-high SES	0.136	0.135	−0.001	0.829
Medium-low SES	0.216	0.201	−0.015	0.023
Low SES	0.195	0.190	−0.012	0.140
Number of siblings	1.510	1.487	−0.032	0.127
Class size	31.75	32.19	0.39	0.196
DNB percentile rank in math	74.17	73.95	0.20	0.699
DNB percentile rank in French	69.31	69.90	0.89	0.122
Test of joint significance	F-stat: 0.983 (p-value: 0.459)			
Predicted undergraduate major				
Major: STEM	0.382	0.384	0.003	0.352
Major: selective STEM	0.175	0.178	0.006	0.081
Major: male-dominated STEM	0.273	0.276	0.004	0.279
N	2,853	2,898	5,751	

Notes: Each row corresponds to a different linear regression with the dependent variable listed on the left, separately for students in Grade 10 (Panel A) and in Grade 12 (Panel B). Columns 1 and 2 show the average value for students in the control and treatment groups, respectively. Column 3 reports the coefficient from the regression of each variable on the treatment group indicator, with the p-value reported in column 4. The regression controls for school fixed effects to account for the fact that randomization was stratified by school, and standard errors are adjusted for clustering at the unit of randomization (class). The F-statistic is from a test of the joint significance of the coefficients in a regression of the treatment group indicator on all student characteristics. High school tracks (Panel A) and undergraduate majors (Panel B) are predicted for each student using the coefficients from a linear regression of the corresponding binary variable (e.g., enrollment in a STEM major) on all student characteristics listed in the table. This model is fitted separately by grade level on the sample of students in the control group.

Table 2 – Female Role Models: Summary Statistics

	All role models (1)	Researchers (Ph.D./ Postdoc) (2)	Professionals (employed by sponsoring firm) (3)
Age (N=51)	33.3	30.0	35.6
	(5.7)	(3.1)	(6.0)
Non-French	0.14	0.10	0.17
holds/prepares for a Ph.D. (N=55)	0.62	1.00	0.38
Graduated from a Grande École	0.39	0.33	0.43
Field: Math, Physics, Engineering	0.23	0.38	0.14
Field: Earth and Life Sciences	0.64	0.62	0.66
Field: Other	0.13	0.00	0.20
Has children (N=52)	0.42	0.19	0.58
Participated in the program the year before	0.25	0.19	0.29
Number of high schools visited	1.8	2.1	1.6
	(0.8)	(0.9)	(0.7)
Number of classroom interventions	5.2	5.9	4.7
	(2.3)	(2.3)	(2.1)
N	56	21	35

Notes: The summary statistics are computed based on information obtained from the L'Oréal Foundation and from the post-intervention survey administered online to collect feedback about the classroom visits. Standard deviations are shown in parentheses below the mean values. Where data are missing for some role models, the number of non-missing values N is indicated in parentheses.

Table 3 – Perceptions of Science-Related Careers

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
Positive perceptions of science-related careers (index)	−0.020	0.245*** (0.028)	0.000	0.023	0.167*** (0.029)	0.000
Science-related jobs require long years of study	0.839	−0.087*** (0.010)	0.000 [0.001]	0.849	−0.074*** (0.010)	0.000 [0.001]
Science-related jobs are monotonous	0.290	−0.032*** (0.012)	0.006 [0.011]	0.318	−0.006 (0.013)	0.633 [0.634]
Science-related jobs are solitary	0.325	−0.061*** (0.012)	0.000 [0.001]	0.303	−0.062*** (0.011)	0.000 [0.001]
Science-related jobs pay higher wages	0.637	0.008 (0.014)	0.535 [0.536]	0.668	0.015 (0.013)	0.237 [0.297]
Hard to maintain work-life balance	0.297	−0.026** (0.012)	0.026 [0.033]	0.283	−0.029** (0.012)	0.014 [0.023]
N		6,475			5,751	
Panel B. Grade 12 (science track)						
Positive perceptions of science-related careers (index)	−0.003	0.312*** (0.034)	0.000	0.003	0.155*** (0.033)	0.000
Science-related jobs require long years of study	0.666	−0.110*** (0.015)	0.000 [0.001]	0.719	−0.091*** (0.014)	0.000 [0.001]
Science-related jobs are monotonous	0.169	−0.019 (0.013)	0.141 [0.141]	0.233	−0.026 (0.016)	0.114 [0.143]
Science-related jobs are solitary	0.228	−0.088*** (0.012)	0.000 [0.001]	0.206	−0.047*** (0.013)	0.000 [0.001]
Science-related jobs pay higher wages	0.531	0.059*** (0.018)	0.001 [0.002]	0.576	0.027* (0.016)	0.093 [0.143]
Hard to maintain work-life balance	0.225	−0.049*** (0.015)	0.001 [0.002]	0.167	−0.012 (0.011)	0.260 [0.260]
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of classroom interventions on students' perceptions of science-related careers, separately by grade level and gender. The sample is restricted to students who completed the post-intervention questionnaire. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4 – Perceptions of Gender Roles in Science

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
More men in science-related jobs	0.628	0.156***	0.000	0.629	0.168***	0.000
		(0.013)	[0.001]		(0.014)	[0.001]
Equal gender aptitude for math (index)	0.115	0.109***	0.000	−0.134	0.148***	0.000
		(0.025)	[0.001]		(0.030)	[0.001]
Women like science less than men	0.157	0.059***	0.000	0.198	0.103***	0.000
		(0.011)	[0.001]		(0.013)	[0.001]
W face discrimination in science-related jobs	0.603	0.127***	0.000	0.527	0.153***	0.000
		(0.013)	[0.001]		(0.014)	[0.001]
N		6,475			5,751	
Panel B. Grade 12 (science track)						
More men in science-related jobs	0.712	0.125***	0.000	0.717	0.149***	0.000
		(0.016)	[0.001]		(0.015)	[0.001]
Equal gender aptitude for math (index)	0.158	0.095***	0.001	−0.161	0.132***	0.001
		(0.028)	[0.001]		(0.040)	[0.002]
Women like science less than men	0.074	0.044***	0.000	0.146	0.073***	0.000
		(0.009)	[0.001]		(0.015)	[0.001]
W face discrimination in science-related jobs	0.624	0.095***	0.000	0.600	0.072***	0.000
		(0.020)	[0.001]		(0.018)	[0.001]
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of classroom interventions on students' perceptions of gender roles in science, separately by grade level and gender. The sample is restricted to students who completed the post-intervention questionnaire. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5 – Stated Preferences and Self-Concept

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value (6)
Panel A. Grade 10						
Taste for science subjects (index)	−0.169	−0.038	0.294	0.197	−0.019	0.533
		(0.036)			(0.031)	
Math self-concept (index)	−0.198	−0.008	0.806	0.231	0.039	0.217
		(0.031)			(0.032)	
Science-related career aspirations (index)	−0.103	0.012	0.695	0.120	0.007	0.801
		(0.030)			(0.029)	
N		6,475			5,751	
Panel B. Grade 12						
Taste for science subjects (index)	−0.002	0.016	0.632	0.002	0.000	0.998
		(0.034)			(0.039)	
Math self-concept (index)	−0.184	0.050	0.202	0.187	0.072**	0.041
		(0.039)			(0.035)	
Science-related career aspirations (index)	−0.045	0.113***	0.002	0.046	0.050	0.131
		(0.037)			(0.033)	
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of classroom interventions on students' taste for science subjects, math self-concept, and science-related career aspirations, separately by grade level and gender. The sample is restricted to students who completed the post-intervention questionnaire. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 6 – Enrollment Status the Following Year

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
All STEM tracks						
Grade 11 : Science track	0.371	−0.008	0.586	0.578	−0.006	0.676
		(0.014)			(0.015)	
General vs. technological STEM track						
Grade 11: Science - general track	0.343	−0.002	0.888	0.436	0.004	0.773
		(0.014)	[0.889]		(0.014)	[0.774]
Grade 11: Science - technological track	0.027	−0.006	0.112	0.141	−0.010	0.235
		(0.004)	[0.224]		(0.009)	[0.470]
N		7,241			6,459	
Panel B. Grade 12 (science track)						
All undergraduate STEM majors						
Major: STEM	0.289	0.024*	0.080	0.470	0.003	0.886
		(0.014)			(0.020)	
Selective vs. non-selective STEM						
Major: selective STEM	0.110	0.035***	0.002	0.232	0.020	0.200
		(0.011)	[0.004]		(0.016)	[0.283]
Major: non-selective STEM	0.178	−0.011	0.322	0.239	−0.017	0.212
		(0.011)	[0.322]		(0.014)	[0.283]
Male- vs. female-dominated STEM						
Major: male-dominated STEM	0.166	0.038***	0.002	0.379	0.017	0.387
(math, physics, computer science)		(0.012)	[0.004]		(0.019)	[0.388]
Major: female-dominated STEM	0.123	−0.015	0.158	0.091	−0.014	0.119
(earth and life sciences)		(0.010)	[0.211]		(0.009)	[0.283]
N		2,827			2,924	

Notes: This table reports estimates of the treatment effects of classroom interventions on students' enrollment outcomes in the academic year following the classroom interventions, i.e. 2016/17, separately by grade level and gender. The enrollment outcomes are measured using student-level administrative data. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 7 – Heterogeneous Treatment Effects on Selective and Male-Dominated STEM Enrollment for Girls in Grade 12: Estimates based on Machine Learning Methods

Panel A. Best linear predictor (BLP) of the CATE $s_0(Z)$ given the ML proxy $S(Z)$				
Parameters:	ATE (β_1)	HET (β_2)		Best ML method
Undergraduate major: selective STEM	0.038	0.762		Elastic Net
p-value	[0.027]	[0.031]		
Undergraduate major: male-dominated STEM	0.036	0.088		Linear model
p-value	[0.064]	[0.731]		
Panel B. Sorted group average treatment effects (GATEs): 20% most and least affected students				
Heterogeneity group:	20% least affected	20% most affected	Difference most−least	Best ML method
Undergraduate major: selective STEM	−0.004	0.139	0.149	Elastic Net
p-value	[1.000]	[0.014]	[0.026]	
Undergraduate major: male-dominated STEM	0.026	0.061	0.038	Elastic Net
p-value	[1.000]	[0.464]	[1.000]	
Panel C. Average characteristics of the 20% most and least affected students (CLAN)				
Heterogeneity group:	20% least affected	20% most affected	Difference most−least	p-value (upper bound)
Enrollment in selective STEM major				
Student characteristics				
Baccalauréat percentile rank in math	17.62	81.39	62.85	0.000
Baccalauréat percentile rank in French	41.45	73.44	32.74	0.000
High SES	0.344	0.637	0.302	0.000
Role model characteristics				
Professional	0.494	0.638	0.148	0.001
Participated in the program the year before	0.141	0.233	0.093	0.015
Non-French	0.133	0.183	0.051	0.228
Has children	0.503	0.417	−0.095	0.064
Age	33.09	32.97	−0.11	1.000
Holds/prepares for a Ph.D.	0.692	0.606	−0.080	0.111
Field: math, physics, engineering	0.316	0.226	−0.099	0.021
Field: earth and life sciences	0.618	0.602	−0.004	1.000
Enrollment in male-dominated major				
Student characteristics				
Baccalauréat percentile rank in math	19.88	79.02	59.45	0.000
Baccalauréat percentile rank in French	41.22	72.10	31.10	0.000
High SES	0.335	0.628	0.296	0.000
Role model characteristics				
Professional	0.530	0.606	0.078	0.170
Participated in the program the year before	0.142	0.240	0.091	0.021
Non-French	0.153	0.164	0.004	1.000
Has children	0.539	0.418	−0.126	0.010
Age	33.15	32.95	−0.17	1.000
Holds/prepares for a Ph.D.	0.705	0.601	−0.103	0.043
Field: math, physics, engineering	0.298	0.237	−0.065	0.186
Field: earth and life sciences	0.657	0.585	−0.075	0.170

Notes: This table reports heterogeneous treatment effects of the program on the undergraduate enrollment outcomes of girls in Grade 12, using the methods developed by Chernozhukov et al. (2018). For each outcome, the conditional average treatment effect (CATE) of role model interventions, $s_0(Z)$, is predicted using five alternative ML methods: Elastic Net, Random Forest, Linear Model, Boosting, and Neural Network. The covariates Z that are used to predict the CATE consist of three indicators for the educational districts of Paris, Créteil, and Versailles, four indicators for students' socioeconomic background (high, medium-high, medium-low, and low), their age, their overall percentile rank in the *Baccalauréat* exam, their percentile ranks in the French and math tests of the exam, and a vector of 56 role model fixed effects. For each outcome, Panel A reports the parameter estimates and p-values (in square brackets) of the Best Linear Predictor (BLP) of the CATE using the best ML method (see Appendix Table M31, Panel A). The coefficients β_1 and β_2 correspond to the average treatment effect (ATE) and heterogeneity loading (HET) parameters in the BLP, respectively. Panel B reports the Sorted Group Average Treatment Effects (GATEs), i.e., the average treatment effects among students in the top and bottom quintiles of the heterogeneous effects induced by the ML proxy predictor $S(Z)$, using the best ML method (see Appendix Table M31, Panel B). Panel C performs a Classification Analysis (CLAN) by comparing the average characteristics of the 20 percent most and least affected students defined in terms of the ML proxy predictor. The parameter estimates and p-values are computed as medians over 100 splits, with nominal levels adjusted to account for the splitting uncertainty. This adjustment implies that the reported p-values should be interpreted as upper bounds for the actual p-values. Further details on the methods are provided in Appendix M.

Table 8 – Heterogeneous Treatment Effects on Student Perceptions: Average Characteristics of the Most and Least Affected Girls in Grade 12

	20% least affected (1)	20% most affected (2)	Difference most−least (3)	p-value (upper bound) (4)
Positive perceptions of science-related careers (index)				
Mean Baccalauréat percentile rank in math	26.62	73.29	46.85	0.000
Class visited by professional	0.483	0.675	0.192	0.000
More men in science-related jobs				
Mean Baccalauréat percentile rank in math	74.87	25.00	−51.03	0.000
Class visited by professional	0.614	0.511	−0.112	0.031
Equal gender aptitude for math (index)				
Mean Baccalauréat percentile rank in math	42.77	50.58	7.89	0.003
Class visited by professional	0.622	0.563	−0.058	0.403
Women like science less than men				
Mean Baccalauréat percentile rank in math	44.47	50.57	5.07	0.090
Class visited by professional	0.592	0.540	−0.035	0.908
Women face discrimination in science-related jobs				
Mean Baccalauréat percentile rank in math	52.15	42.79	−8.81	0.001
Class visited by professional	0.568	0.570	0.011	1.000
Taste for science subjects (index)				
Mean Baccalauréat percentile rank in math	41.36	54.71	13.63	0.000
Class visited by professional	0.436	0.678	0.227	0.000
Math self-concept (index)				
Mean Baccalauréat percentile rank in math	52.22	42.10	−10.65	0.000
Class visited by professional	0.512	0.582	0.071	0.240
Science-related career aspirations (index)				
Mean Baccalauréat percentile rank in math	44.70	47.78	2.36	0.712
Class visited by professional	0.375	0.762	0.389	0.000

Notes: This table reports the average characteristics of Grade 12 girls in the top and bottom quintile of predicted treatment effects on student perceptions, using the methods developed by Chernozhukov et al. (2018). For each outcome, the conditional average treatment effect (CATE) of role model interventions, $s_0(Z)$, is predicted using five alternative ML methods: Elastic Net, Random Forest, Linear Model, Boosting, and Neural Network. The covariates Z that are used to predict the CATE consist of three indicators for the educational districts of Paris, Créteil, and Versailles, four indicators for students' socioeconomic background (high, medium-high, medium-low, and low), their age, their overall percentile rank in the *Baccalauréat* exam, their percentile ranks in the French and math tests of the exam, and a vector of 56 role model fixed effects. For each outcome, the table compares the average characteristics of the students in the top and bottom quintile of treatment effects, as predicted by the best ML proxy predictor based on the Group average treatment effects (GATEs) targeting of the CATE (see Appendix Table M31, Panel B). The characteristics reported in this table are the students' average percentile rank in math (in the *Baccalauréat* exams) and the share exposed to a role model with a professional rather a research background. The parameter estimates and p-values are computed as medians over 100 splits, with nominal levels adjusted to account for the splitting uncertainty. This adjustment implies that the reported p-values should be interpreted as upper bounds for the actual p-values. This adjustment implies that the reported p-values should be interpreted as upper bounds for the actual p-values. The average treatment effects among the 20 percent most and least affected students can be found in Panel B of Appendix Table M32. Further details on the methods are provided in Appendix M.

Table 9 – Correlation between Conditional Average Treatment Effects (CATEs) for Girls in Grade 12

	Bivariate correlation with the CATE on enrollment in a selective STEM program	
	Estimate ($\hat{\rho}_{A,B\|Z}$)	95% confidence interval
Conditional average treatment effect (CATE) on:		
Positive perception of science-related careers (index)	0.96	[0.21, 5.30]
More men in science-related jobs	−0.68	[−3.23, −0.01]
Equal gender aptitude for math (index)	0.19	[−1.24, 2.05]
Women like science less than men	0.21	[−1.43, 3.23]
Women face discrimination in science-related jobs	−0.34	[−2.22, 0.56]
Taste for science subjects (index)	0.71	[0.04, 3.96]
Math self-concept (index)	−0.07	[−1.84, 1.40]
Science-related career aspirations (index)	0.36	[−0.51, 2.01]

Notes: This table reports, for girls in Grade 12, estimates of the bivariate correlation $\rho_{A,B|Z}$ between the Conditional Average Treatment Effect (CATE) on enrollment in a selective STEM program, denoted by $s_0^B(Z)$, and the CATE on each of the potential mediators listed in the table, denoted by $s_0^A(Z)$. The proxy predictor of the CATE on selective STEM enrollment, denoted by $S^B(Z)$, is estimated using the Elastic Net method, as it has the best performance based on the Best Linear Predictor (BLP) targeting of the CATE for this outcome. The proxy predictor of the CATE on the potential mediator Y^A, denoted by $S^A(Z)$, is estimated using the ML method that has the best performance based on the BLP targeting of the CATE on the corresponding outcome (see Appendix Table M31, Panel A). An indication of the quality of these predictions is provided by the heterogeneity loading (HET) parameter of the BLP (see Appendix Table M32, Panel A). For each random split of the data, the correlation coefficient $\rho_{A,B|Z}$ is estimated as $\hat{\rho}_{A,B|Z} = \text{Sign}(\hat{\beta}_2^{A|B})\sqrt{\hat{\beta}_2^{A|B}\hat{\beta}_2^{B|A}}/\sqrt{\hat{\beta}_2^{A|A}}\sqrt{\hat{\beta}_2^{B|B}}$, where $\hat{\beta}_2^{k|l}$ is the estimated heterogeneity loading parameter of the BLP of $s_0^k(Z)$ based on $S^l(Z)$ (with $k, l \in \{A, B\}$), using the methods in Chernozhukov et al. (2018). The covariates Z that are used to predict the CATEs consist of three indicators for the educational districts of Paris, Créteil, and Versailles, four indicators for students' socioeconomic background (high, medium-high, medium-low, and low), their age, their overall percentile rank in the *Baccalauréat* exam, their percentile ranks in the French and math tests of the exam, and a vector of 56 role model fixed effects. For each pair of outcomes, columns 1 and 2 report the estimated correlation between the CATEs and its 95 percent confidence interval, respectively. Estimates and confidence intervals are computed as medians over the first 100 random data splits for which $\hat{\rho}_{A,B|Z}$ can be computed. For each data split, the confidence intervals are obtained using a clustered bootstrap procedure. The nominal level of the median of confidence intervals is adjusted to account for the splitting uncertainty, using the method of Chernozhukov et al. (2018). This adjustment implies that the reported confidence intervals should be interpreted as lower and upper bounds for the true lower and upper limits of the confidence intervals. Further details on the methods are provided in Appendix M.

(For Online Publication)

Appendix to

Do female Role Models Reduce the Gender Gap in Science? Evidence from French High Schools

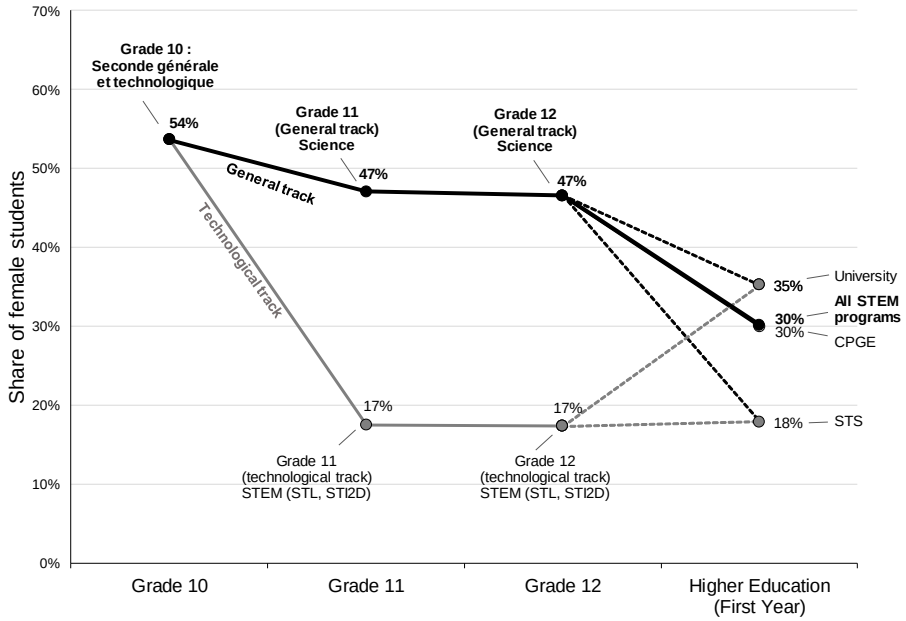
Thomas Breda
Julien Grenet
Marion Monnet
Clémentine Van Effenterre

April 2020

List of Appendices

A Female Representation in STEM Studies in France

(a) Academic year 2016/17



(b) Academic year 2006/07

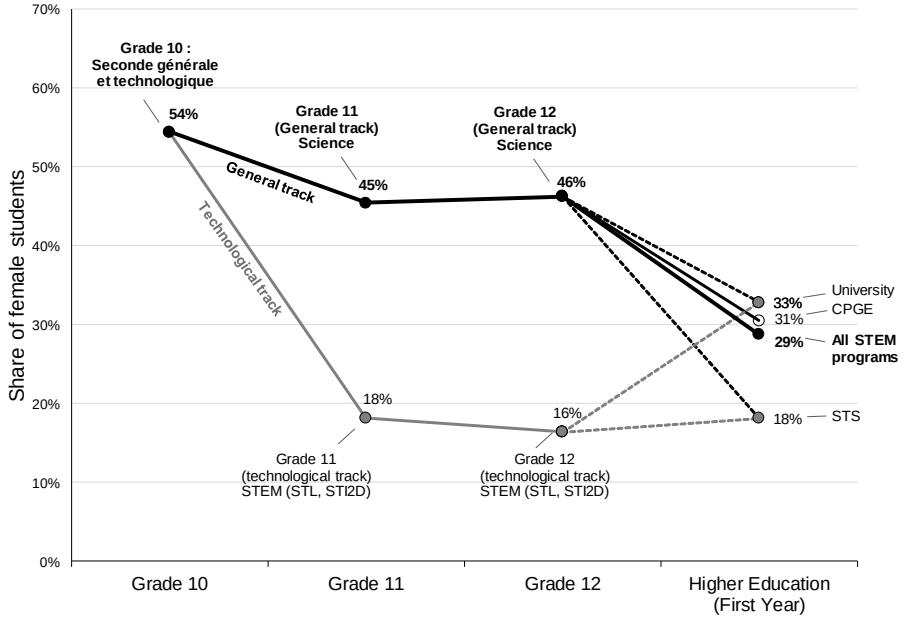
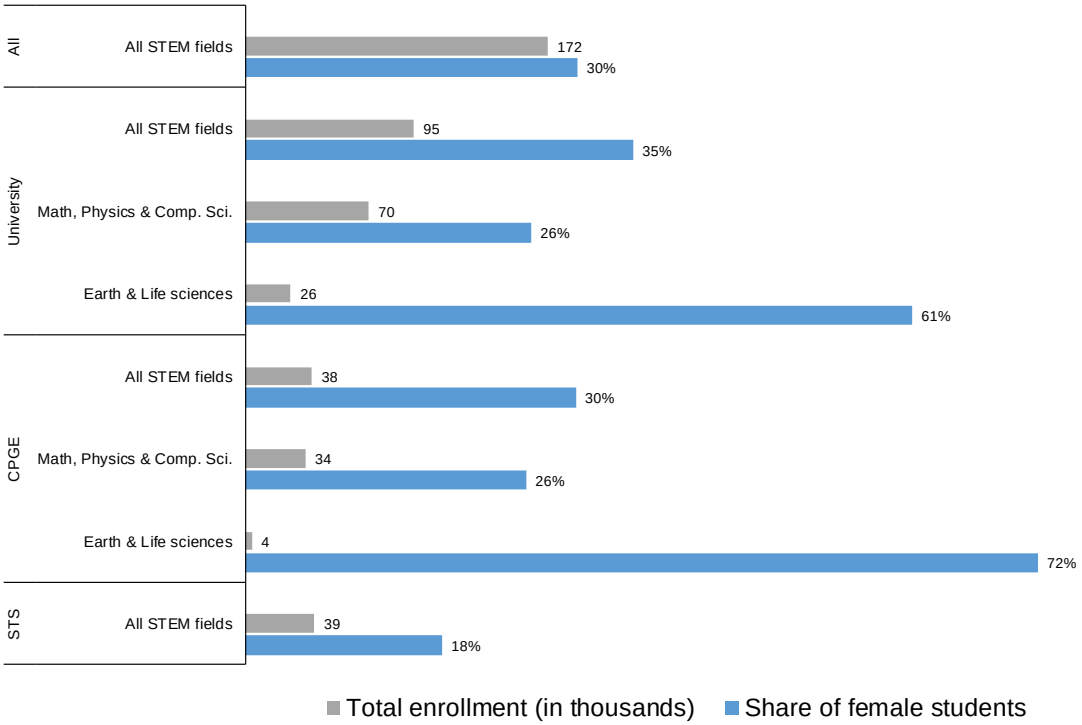


Figure A1 – Female Underrepresentation in STEM Fields of Study in France

Notes: The figure shows the share of female students in the first year of high school (Grade 10), in the STEM-oriented high school tracks in Grade 11 and Grade 12, and in the STEM fields of first-year undergraduate programs which are either selective (CPGE and STS) or non-selective (University). After completing Grade 10, high school students are directed either to the general track (leading to the general *Baccalauréat*) or to the technological track (leading to the general *Baccalauréat*). In the general track, the Science (S) sub-track specializes in STEM fields. In the technological track, the two STEM-oriented sub-tracks are STL and STI2D.
Sources: Authors' calculations from the SCOLARITE (MENJ-DEPP) and SISE (MESRI-SIES) datasets for 2006/07 (Panel b) and 2016/17 (Panel a).

(a) STEM fields of study



(b) Non-STEM fields of study

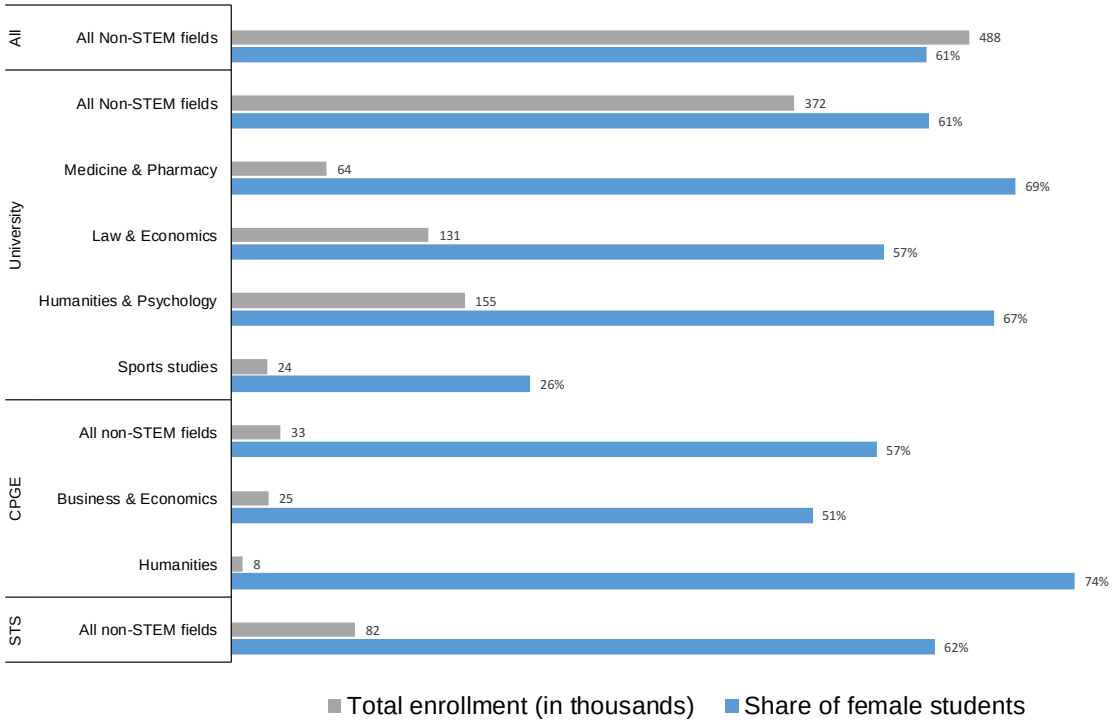


Figure A2 – First-Year Undergraduate Students: Total Enrollment and Share of Female Students in STEM and non-STEM Fields of Study, Academic Year 2016/17

Sources: Authors' calculations from the SCOLARITE and SISE datasets (MENJ-DEPP and MESRI-SIES) for 2016/17.

B Gender Pay Gap Among College Graduates in France

This appendix provides descriptive evidence on the entry-level gender pay gap among French college graduates holding a master's degree and analyzes the contribution of gender segregation in college majors to this gap. The objective of this analysis is to better understand whether the effects of the role model interventions on female students' choice of studies can be expected to reduce the gender pay gap. Section B.1 describes the data sources, while Section B.2 discusses the empirical results.

B.1 Data

Unfortunately, we cannot rely exclusively on administrative data to provide empirical evidence on the gender pay gap by field of study in France, as it is currently not possible to link administrative data on students enrolled in higher education with administrative data on wages and income tax returns. Instead, our analysis of the gender pay gap among college graduates is based on the combination of aggregate statistics on student enrollment by college major and gender with survey information on the starting wages of recent cohorts of college graduates.

Data sources. In France, gender segregation and gender pay gaps by college major can be analyzed for the population of college graduates who obtained their master's degree (or equivalent) in 2015 or 2016. For this purpose, we combine several administrative and survey data sources.

SISE Résultats 2015. this individual-level administrative dataset covers all students enrolled in public universities during the academic year 2015/16 and provides detailed information on each student's degree program and field of study.

Enquête Professionnelle des Diplômés de Master 2015 (EPDM). This survey was conducted in December 2017 by the Ministry of Higher Education to collect information on the transition of master's graduates to the labor market. The survey was targeted at students who obtained their master's degree in 2015 and who entered the labor market within one year after graduation, with an overall response rate of 70 percent. As part of this survey, master's graduates were asked to report their annual earnings 18 months after graduation. Our analyses are based on the survey's public use files, which provide aggregate statistics by gender and college major.[A.1]

Enquête sur l'Insertion des Diplômés des Grandes Écoles 2018 (EIDGE). This survey was conducted in 2018 by the Conférence des Grandes Écoles (CGE), a not-for-profit association representing French elite graduate schools. The *Grandes Écoles*, which award a diploma equivalent to a master's degree, recruit their students through highly competitive national exams taking place at the end of two-year undergraduate selective STEM and non-STEM preparatory courses (*Classes Préparatoires aux Grandes Écoles* or CPGE). The survey was targeted at students who graduated between 2015 and 2017 from one of the 184 *Grandes Écoles* that are members of the CGE, with an overall response rate of 48 percent. Our analyses are based on the aggregate statistics published by the CGE separately by gender and by type of *Grande École* (i.e., engineering schools, business schools, and other schools).[A.2] We only consider students who graduated from a *Grande École* in 2016, since annual earnings 24 months after graduation are only available for this cohort.

Grouping of college majors. The above data sources can be combined to compute the number of female and male master's students who graduated from university in 2015 or from a

[A.1] `https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-insertion_professionnelle-master_donnees_nationales/information/` (accessed on August 2, 2019).

[A.2] `https://www.cge.asso.fr/themencode-pdf-viewer/?file=https://www.cge.asso.fr/wp-content/uploads/2018/06/2018-06-19-Rapport-2018.pdf` (accessed on August 2, 2019).

Grande École in 2016, separately by college major.

The Ministry of Higher Education's official classification comprises 54 college majors. For the purpose of our analysis, we group these college majors into the following broad categories:

- Non-STEM majors (35 in total): this category includes master's degree programs in law, economics, management, humanities, psychology, social sciences, medicine, pharmacy, sports studies as well as degrees from non-STEM *Grande Écoles* (e.g., business schools, schools of journalism, schools of architecture).
- STEM majors (19 in total): this category includes master's degree programs in STEM fields as well as degrees from engineering schools (*Grande Écoles d'ingénieurs*).
- Among STEM majors, we distinguish between engineering schools (all of which are selective and are classified as a single major) and non-selective STEM master's degrees at university (18 in total).
- Among non-selective STEM majors, we further distinguish between male-dominated majors (16 in total) and female-dominated majors (2 in total: chemistry and earth and life sciences), based on whether the share of female students among master's graduates in the corresponding field of study is below or above 50 percent. This distinction does not apply to selective STEM majors, since almost all engineering schools are male-dominated.

Earnings information. The EPDM and EIDGE surveys provide information on graduates' average median gross salary (*salaire brut annuel médian*) separately by gender and college major. Starting wages are measured 18 months after graduation for master's graduates and 24 months after graduation for *Grandes Écoles* graduates. Note that since we do not have access to the individual-level survey data, median earnings by broad categories of college majors can only be approximated as the average of the median earnings in each of the majors that form these broad categories.

B.2 College Majors and the Gender Pay Gap

Combining the above data sources, we provide descriptive evidence on the median starting wages of female and male graduates across the broad categories of college majors. We then analyze the contribution of gender segregation in college majors to the overall entry-level gender pay gap.

Gender composition of STEM and non-STEM majors. The first three columns of Table B1 show the distribution of master's-level graduates across the broad categories of college majors that we defined above, along with the share of female graduates in each category. The summary statistics indicate that while female students represent 52 percent of master's level graduates, they are strongly underrepresented in STEM majors (34 percent). Female underrepresentation is more pronounced in selective (male-dominated) STEM majors (female share: 30 percent) than in non-selective STEM majors (female share: 40 percent). Among non-selective STEM majors, female students represent only 29 percent of graduates in male-dominated fields such as mathematics, physics, or computer science, compared to 60 percent of graduates in female-dominated fields such as chemistry and earth and life sciences.

Starting wages of STEM and non-STEM graduates. The comparison of starting wages by broad college major category confirms that female graduates tend to be overrepresented in lower-paying majors (see columns 3–5 of Table B1). Female graduates holding a STEM degree have a median starting wage of 29,984 euros, which is 7.4 percent higher than the median starting wage of female graduates holding a non-STEM degree (27,913 euros). Strikingly, the

wage premium for female graduates in STEM appears to be almost entirely driven by selective (male-dominated) STEM degrees (16.4 percent). By contrast, the wage premium attached to non-selective STEM degrees is close to zero (−0.5 percent). The low apparent return to non-selective STEM degrees masks substantially different returns between male-dominated and female-dominated majors: while the wage premium attached to male-dominated non-selective STEM majors is of 4.2 percent for female graduates compared to non-STEM majors, a wage penalty of 4.7 percent is attached to female-dominated non-selective STEM majors.

Female underrepresentation in STEM: contribution to the gender pay gap. The last three columns of Table B1 indicate that across all categories of programs, male graduates earn a median annual starting wage of 32,122 euros, compared to 28,411 euros for female graduates. This amounts to an overall gender pay gap of 3,711 euros per year, or 11.6 percent of male pay.

Although the overrepresentation of female graduates in lower-paying non-STEM and female-dominated STEM majors is a likely contributor to the overall gender pay gap, it is clearly not the sole cause, as gender differences in median earnings are observed within each broad category of college majors. Interestingly, however, the gender wage gap is lower in each category of STEM majors than in non-STEM majors. This finding is consistent with similar evidence for the U.S. (Beede et al., 2011).

To shed light on the contribution of gender segregation in fields of study to the overall entry-level gender pay gap, we adopt a method similar to that used by McDonald and Thornton (2007) in estimating what the overall female-male starting wage gap would be if female graduates had the same distribution of college majors as male graduates.

Since our interest is in measuring the specific contribution of the different dimensions of female underrepresentation in STEM majors (STEM vs. non-STEM, selective vs. non-selective STEM, male-dominated vs. female-dominated non-selective STEM), we construct counterfactual wage gaps by considering increasingly disaggregated groups of majors.

We start by estimating the counterfactual wage gap if female graduates had the same distribution of STEM vs. non-STEM majors as male graduates, while keeping fixed females' marginal distribution of majors within each of these two broad categories. Put differently, we apply female median earnings in STEM vs. non-STEM degrees to the male distribution of graduates in both categories of majors to recalculate the overall gender pay gap. This counterfactual wage gap, which we denote by $\tilde{\Delta}_w$, is constructed as follows:

$$\tilde{\Delta}_w = 1 - \frac{(\bar{w}_s^f N_s^m + \bar{w}_{ns}^f N_{ns}^m)}{(\bar{w}_s^m N_s^m + \bar{w}_{ns}^m N_{ns}^m)},$$

where $\bar{w}_k^g$ and N_k^g denote the median earnings and the number of graduates of gender g (m: males; f: females) in college major category k (s: STEM; ns: non-STEM), respectively. The contribution of female underrepresentation in STEM programs to the gender pay gap is then measured as $\Delta_w - \tilde{\Delta}_w$, where Δ_w denotes the observed overall pay gap between male and female graduates.

To measure the contribution of gender segregation between selective and non-selective STEM majors, we construct a second counterfactual wage gap in a similar manner, except that college majors are now grouped into three categories: non-STEM, selective STEM, and non-selective STEM. To measure the contribution of gender segregation between male-dominated and female-dominated STEM majors, we repeat this exercise after grouping college majors into four categories: non-STEM, selective STEM, non-selective male-dominated STEM, and non-selective female-dominated STEM. The contribution of gender segregation between majors within both male- and female-dominated non-selective STEM is measured by ungrouping all STEM majors. Finally, we ungroup all non-STEM majors to evaluate the contribution of gender

segregation between non-STEM majors. The corresponding counterfactual measures what the overall gender gap would be if women had the same distribution as men across all 54 STEM and non-STEM college majors.

Results. The results of this decomposition exercise are shown in Table B2 along with the observed gender pay gap. The contributions of gender segregation between the different categories of college majors to the gender pay gap are reported in column 1 and are expressed as percentages of the total in column 2. We find that the gender imbalances across all college majors "explain" 40 percent of the gender pay gap among college graduates. Two-thirds of this explained part (26.5 percent of the total wage gap) can be attributed to the unequal representation of female and male graduates in STEM vs. non-STEM majors, on the one hand, and between the different majors within STEM, on the other hand. The remain third of the explained part of the gap (13.4 percent of the total) is due to gender segregation between non-STEM majors, the lowest-paying majors (humanities) being typically more female-dominated (77 percent) than the highest-paying ones (law and economics, in which the female share is 59 percent).

The 26.5 percent STEM-related gender pay gap can be decomposed as follows. Increasing the share of female graduates holding a STEM degree to that of males without changing females' marginal distribution of STEM majors is associated with a 14.0 percent reduction in the gender pay gap. In line with the evidence from Table B1, further reassigning female graduates from non-selective STEM majors to (male-dominated) selective STEM majors in order to match the relative shares of selective and non-selective STEM majors among male graduates would reduce the gender gap by an additional 6.5 percent from the baseline. Finally, reassigning female graduates from non-selective female-dominated STEM majors to non-selective male-dominated STEM majors would trigger an extra 4.3 percent reduction in the gender pay gap, while further reassigning female students between majors within male- and female-dominated programs would result in an extra 1.8 percent reduction from the baseline.

Altogether, these findings suggest that the underrepresentation of female students in STEM majors accounts for approximately 25 percent of the entry-level gender pay gap among college graduates in France. Moreover, almost half of this STEM-related gender pay gap can be attributed to the fact that within STEM majors, female graduates are relatively less likely than males to be enrolled in those with the largest wage premium, i.e., the selective and male-dominated STEM majors.

Table B1 – Starting Wage Among College Graduates Holding a Master's Degree or Equivalent, Classes of 2015/16

	Graduates: classes of 2015/16			Wage 18/24 months after graduation (survey)				
				Female graduates		Male graduates		
	Number of graduates	% of total	Female share (%)	Median wage (euros)	Relative Median wage (non-STEM majors: 100)	Median wage (euros)	Relative Median wage (non-STEM majors: 100)	Gender pay gap (%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All majors (54)	166,600	100.0	51.5	28,411	-	32,122	-	11.6
Non-STEM majors (35)	106,997	64.2	61.1	27,913	100.0	31,302	100.0	10.8
STEM majors (19)	59,603	35.8	34.3	29,984	107.4	32,972	105.3	9.1
of which:								
Selective (male-dominated) STEM majors (Engineering schools)	31,463	18.9	29.7	32,500	116.4	34,800	111.2	6.6
Non-Selective STEM majors (18)	28,140	16.9	39.6	27,767	99.5	30,530	97.5	9.1
of which:								
Male-dominated majors (16)	18,874	11.3	29.4	29,077	104.2	31,371	100.2	7.3
Female-dominated majors (2)	9,266	5.6	60.3	26,596	95.3	27,581	88.1	3.6

Notes: This table reports summary statistics on gender segregation and gender pay gaps for the population of college graduates who obtained their master's degree (or equivalent) in 2015 or 2016. The 54 college majors are grouped into two broad categories: non-STEM majors (master's degrees in economics, management, humanities, psychology, social sciences, sports studies, medicine, pharmacy, and non-STEM *Grandes Écoles* such as business schools or schools of journalism) and STEM majors (master's degrees in STEM fields and degrees from engineering schools); STEM majors are further broken down between selective (engineering schools) and non-selective majors (master's degree at university); among non-selective majors, we further distinguish between male-dominated and female-dominated majors, based on whether the share of female graduates in the corresponding field of study is below or above 50 percent. Column 1 shows the number of graduates per broad category of college majors using the administrative dataset SISE 2015/16 (for university graduates who obtained their master's degree in 2016) and the EIDGE survey (for students who graduated from *Grandes Écoles* in 2016). Median gross annual wages (columns 4 and 6) are computed from aggregate statistics by gender and college major from the EPDM and EIDGE surveys. Entry-level wages are measured 18 months after graduation for master's graduates and 24 months after graduation for *Grandes Écoles* graduates. Median wages by broad categories of college majors are approximated as the average of the median wages in each of the majors that form these broad categories.
Sources: Columns 1–3: SISE 2015/16 and Enquête sur l'Insertion des Diplômés des Grandes Écoles 2018 (EIDGE); columns 4–8: Enquête Professionnelle des Diplômés de Master 2015 (EPDM) and EIDGE.

Table B2 – Contribution of Gender Segregation in College Majors to the Entry-Level Gender Wage Gap Among College Graduates, Classes of 2015/16

	Gender pay gap (relative to male pay) (1)	Share of the gender wage gap (2)
Total wage gap	0.116	100.0%
Contribution of gender segregation in college majors to the wage gap:		
Explained by unequal gender distribution between majors	0.046	40.0%
of which:		
between STEM/non-STEM majors and between majors within STEM	0.031	26.5%
of which:		
between STEM and non-STEM majors	0.016	14.0%
between selective and non-selective STEM majors	0.007	6.5%
between male- and female-dominated non-selective STEM majors	0.005	4.3%
between majors within male- and female-dominated non-selective STEM	0.002	1.8%
between majors within non-STEM	0.016	13.4%
Unexplained by unequal gender distribution between majors	0.069	60.0%

Notes: This table provides a decomposition of the total entry-level wage gap between male and female college graduates who obtained their master's degree or equivalent in 2015 (university graduates) or in 2016 (*Grandes Écoles* graduates). Entry-level wages are measured as median annual gross wages by gender and college majors, 18 months after graduation for master's graduates, and 24 months after graduation for *Grandes Écoles* graduates. To measure the contribution of the unequal gender representation across college majors, counterfactual wage gaps are constructed using increasingly disaggregated groups of college majors. The contribution of gender segregation between STEM and non-STEM majors is measured as the observed gender wage gap minus the counterfactual wage gap that would be observed if female graduates had the same distribution of STEM and non-STEM majors as male graduates, while keeping fixed females' marginal distribution of majors within each of these two broad categories. The contribution of gender segregation between selective and non-selective STEM majors is estimated in a similar manner, except that the counterfactual gender wage gap is estimated by reassigning female graduates from non-selective STEM majors to selective STEM majors to match the relative shares of selective and non-selective STEM majors among male graduates. The other components of the gender wage gap are measured by sequentially ungrouping college majors to compute counterfactual gender wage gaps. The contributions of gender segregation between the different categories of college majors to the gender wage gap are shown in column 1 and are expressed as percentages of the total in column 2.
Sources: See notes of Table B1.

C Program Details

(a) First Video: "Jobs in Science: Beliefs or Reality?"

(b) Second Video: "Are we All Equal in Science?"

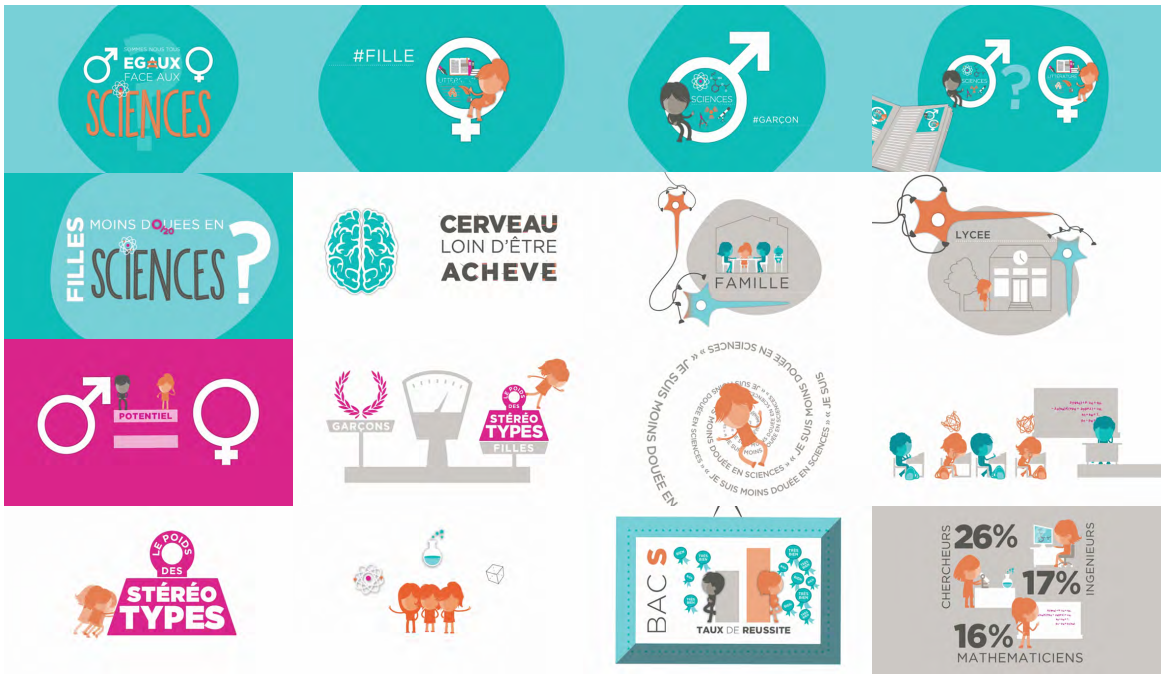


Figure C3 – Screenshots of the Two Videos Shown During the Role Model Interventions

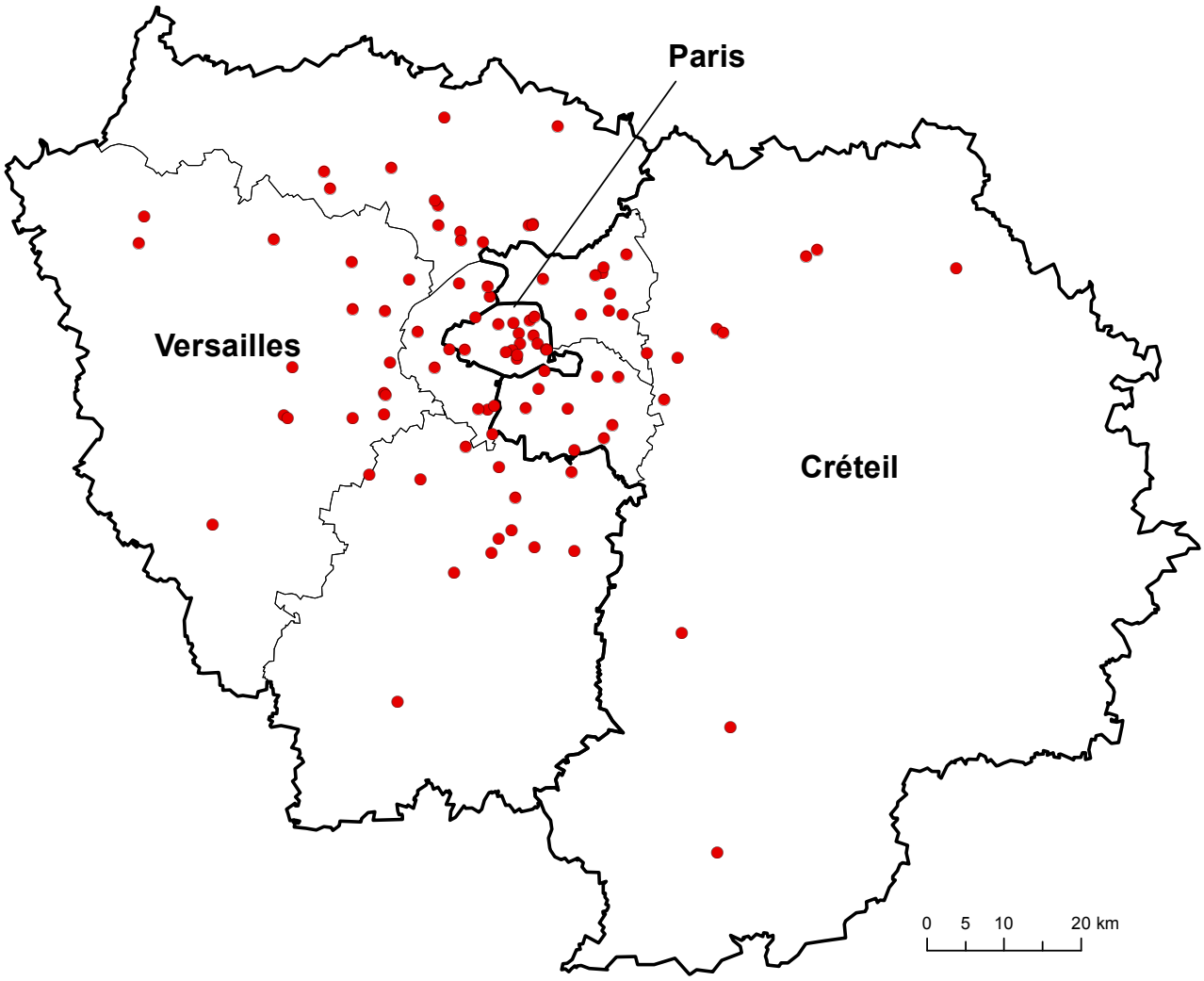


Figure C4 – Participating High Schools

Notes: The thick lines represent the boundaries of the three education districts (*académies*) of the Paris region (Paris, Créteil and Versailles). The solid circles show the location of the 98 high schools that participated in the program evaluation.

D Survey Instruments

D.1 Role Model Survey

The female role models conducting the classroom interventions were invited to complete the following online survey after each visit to a school. Each school visit typically consisted of three classroom interventions (two in Grade 10 and one in Grade 12 (science track)).

Q1. Please indicate your name and surname
Name: ..... Surname: .....

Q2. Please select the high school that you visited
[Drop down menu with the list of participating schools]

Q3. On which date did your classroom interventions take place?
Day ☐☐ Month ☐☐ Year ☐☐☐☐

Q4. Were the classes you visited those planned in the schedule?
- First intervention: ☐ Yes ☐ No
- Second intervention: ☐ Yes ☐ No
- Third intervention: ☐ Yes ☐ No

[The role models were asked to answer questions 5 to 20 for each classroom intervention]

Q5. Please enter the name/identifier of the class
[Free text field]

Q6. At what time did the classroom intervention start?
Starting time: ☐☐:☐☐ [AM/PM]

Q7. Was the teacher present?
☐ Yes ☐ No

Q8. [If "Yes" to Q7] The teacher was
☐ A man ☐ A woman

Q9. [If "Yes'' to Q7] What was the subject taught by the teacher?
☐ Earth and life sciences
☐ Mathematics
☐ English
☐ French
☐ Physical and sports education
☐ History and geography
☐ Physics and chemistry
☐ Other (Specify: .....)
☐ I don't know

Q10. [If "Yes" to Q7], did he/she seem interested?
☐ Yes ☐ Rather yes ☐ Rather no ☐ No

Q11. Apart from the teacher, was there another adult in the classroom (e.g., educational adviser, another teacher)?
☐ Yes ☐ No

Q12. Could you display the PowerPoint presentation?
☐ Yes ☐ No

Q13. Could you show the videos?
☐ Yes ☐ No

Q14. Overall, you found that the students were...
[Possible answers: Yes, very much/Yes, somewhat/Not really/Not at all/Mixed]
- interested
- engaged in the discussion
- inattentive or difficult to contain

Q15. Did you run into logistical problems (e.g., the teacher was not informed of the visit)?
☐ Yes (Specify: .....) ☐ No

Q16. Would you say that gender stereotypes (e.g., "science is not for girls") were strong among the students?
☐ Yes, very much ☐ Rather yes ☐ Rather no ☐ Not at all

Q17. Was there a discipline problem that required an interruption?
☐ Yes ☐ No

Q18. Overall, how did your classroom intervention go?
[Possible answers: Very well/Well/Average/Not so well/Not well at all]

Q19. Do you feel that your intervention was well suited to the students?
☐ Yes, very much ☐ Rather yes ☐ Rather no ☐ Not at all

Q20. Based on the class's reactions (questions, smiles, discussion with students at the end of the intervention...), how receptive do you think the students were to the following messages?
[Possible answers: This topic was not addressed/This topic was addressed and the students were very receptive/This topic was addressed and the students were rather receptive/This topic was addressed but the students were not very receptive/This topic was addressed but the students were not at all receptive]
- "Science is everywhere"
- "Jobs in science are fulfilling"
- "Jobs in science are for girls too"
- "Jobs in science pay well"
- Short videos

Q21. Do you have any comments or suggestions? (please feel free to expand)
[Free text field]

D.2 Student Survey

Survey Preamble

You are about to answer a questionnaire on students' attitudes towards science and science-related careers.

This study will not influence decisions regarding your education in any way. It is conducted independently by a team of researchers from the Paris School of Economics, a public research institution that is affiliated with the Centre National de la Recherche Scientifique (CNRS), with the aim of better understanding high school students' educational and career choices. The study's methodology is based on the combined use of a student survey and of administrative data collected by the statistical offices of the education districts of the Île-de-France region.

We are interested in knowing your views. We thank you in advance for answering all the questions even if they sound similar. If none of the response categories fits your answer exactly, please choose the one that is closest to your opinion.

Your answers will remain completely confidential. A serial number has been generated so ensure that the researchers never have access to your name and surname. No personal information collected through this survey will be communicated to the regional education authority, to your school, to your teachers, or to your family. To protect the confidentiality of your answers, please make sure that you insert your survey questionnaire into the sealed envelope that was provided to you before you hand it to your teacher.

We thank you in advance for taking the time to answer our questions. Your participation is very important to our understanding of students' decision making.

General Questions

G1. What is today's date?
Day ☐☐ Month ☐☐ Year ☐☐☐☐

G2. You are:
☐ A girl ☐ A boy

G3. What is your date of birth?
Day ☐☐ Month ☐☐ Year ☐☐☐☐

G4. Do you have older brothers or sisters?
Older brother(s): ☐ Yes ☐ No
Older sister(s): ☐ Yes ☐ No

G5. Is your father's occupation related to science?
☐ Yes ☐ No ☐ I don't know

G6. Is your mother's occupation related to science?
☐ Yes ☐ No ☐ I don't know

G7. How often do you participate in the following activities? (check only one box per row)
[Possible answers: Once a day/Once a week/Once a month/Never]

- Play video games
- Play a team sport
- Play board games or strategy games
- Participate in sports competitions
- Watch TV shows about science (e.g., "E=M6")
- Read comic books
- Go on Facebook

- Go out with friends
- Spend time with your family outside meal times

Part A: Subjects Studied

A1. On a scale from 0 ("not at all") to 10 ("very much"), how much do you enjoy the following subjects? (check only one box per row) [10 boxes from 0 to 10]
[The order of subjects was randomized across respondents]
- Earth and life sciences
- Mathematics
- English
- French [Grade 10 students only]
- Philosophy [Grade 12 students only]
- Physical and sports education
- History and geography
- Physics and chemistry

A2. How would you assess your level in the following subjects? (check only one box per row)
"My level in this subject is..." [Very weak/Weak/Average/Good/Very Good]
[The order of subjects was randomized across respondents]
- Earth and life sciences
- Mathematics
- English
- French [Grade 10 students only]
- Philosophy [Grade 12 students only]
- Physical and sports education
- History and geography
- Physics and chemistry

A3. How would you assess your level in the following subjects, compared to the average of the boys in your class? (check only one box per row)
"Compared to the average of the boys in my class, I would say that my level in this subject is..." [Much worse/Somewhat worse/Equal/Somewhat better/Much better]
- French [Grade 10 students only]
- Mathematics
- Earth and life sciences [Grade 12 students only]

A4. How would you assess your level in the following subjects, compared to the average of the girls in your class? (check only one box per row)
"Compared to the average of the girls in my class, I would say that my level in this subject is..." [Much worse/Somewhat worse/Equal/Somewhat better/Much better]
- Mathematics
- French [Grade 10 students only]
- Earth and life sciences [Grade 12 students only]

[Note: The order of questions A3 and A4 was randomized across respondents]

A5. To what extent do you agree with the following statements?
[Possible answers: Strongly agree/Somewhat agree/Disagree/Strongly disagree]
- I feel lost when I try to solve a math problem
- I often worry that I will struggle in math class
- If I make enough effort, I can do well in science subjects

Part B: Choice of Studies

B1. [Grade 10 students only] Which high school track would you like to pursue next year? (multiple answers are possible)

- ☐ Grade 11 - Science (*Première S*)
- ☐ Grade 11 - Humanities (*Première L*)
- ☐ Grade 11 - Social Sciences (*Première ES*)
- ☐ Grade 11 - Technological track (*Première Technologique*)
  If so, which one? ☐ STI2D ☐ STD2A ☐ STMG ☐ ST2S ☐ STL ☐ S2TMD ☐ STHR ☐ STAV
- ☐ Grade 11 - Vocational track (*Première Professionnelle*)
- ☐ Other (Specify: .....)
- ☐ I don't know

[Grade 12 students only] What studies would you like to pursue after high school? (multiple answers are possible)

- ☐ University
- ☐ Classes Préparatoires aux Grandes Écoles (standard or integrated)
- ☐ STS (technician's diploma)
- ☐ IUT (university institutes of technology)
- ☐ Specialized schools (paramedical and social care, architecture, journalism, public affairs, arts, etc.)
- ☐ Other (Specify: .....)
- ☐ I don't know

B2. When did you finalize this choice? (check only one box)

- ☐ Before September 2015
- ☐ In September 2015
- ☐ In October 2015
- ☐ In November 2015
- ☐ In December 2015
- ☐ In January 2016
- ☐ In February 2016
- ☐ In March 2016
- ☐ I haven't decided yet

B3. Are you still unsure about your choice of studies?
☐ Yes, a lot ☐ Yes, a little ☐ No

B4. If you are planning to pursue higher education, which field are you considering? (multiple answers are possible)

- ☐ Earth and life sciences
- ☐ Mathematics, Physics, Computer science, Engineering
- ☐ Health
- ☐ Paramedical and social care
- ☐ Law, Economy, Management, Marketing, Communications
- ☐ Literature, History, Geography, Psychology, Sociology, Philosophy, Linguistics
- ☐ Sports studies
- ☐ Arts
- ☐ Other (Specify: .....)
- ☐ I don't know

B5. Are your parents pushing you to pursue scientific studies?
☐ Yes, a lot ☐ Yes, a little ☐ Not at all

B6. List two jobs you would like to do later in life
- Job 1: [Fill in by hand]
- Job 2: [Fill in by hand]

B7. Which of the following jobs might you be interested in as a career? (check only one box per row)
"I would like this job..." [Yes/Rather yes/Rather no /No]
[The order of the jobs was randomized across respondents]
- Engineer
- Computer scientist
- Pharmacist
- Chemist
- Psychologist
- Physician
- Researcher in biology
- Renewable energy technician
- Industrial designer
- Lawyer

B8. Below is a list of factors that might influence your choice of studies. On a scale of 0 to 10, indicate how important each factor is to you (0: "this factor is not important at all in my choice of studies"; 10: "this factor is essential in my choice of studies"). Check only one box per row (0 = not important at all / 10 = essential)
- Interest in the field of study
- The opportunity to specialize quickly
- Being able to access many jobs
- The risk that other study programs might be too difficult
- The ease of finding a stable job
- Career and salary prospects
- The sense of fit
- The workload
- Being surrounded by girls
- Being surround by boys

Part C: Attitudes towards Science

C1. What is your opinion regarding the following statements?
[Possible answers: Strongly agree/Somewhat agree/Disagree/Strongly disagree]
- I like science in general
- There are jobs in science that I find interesting
- I could see myself working in a science-related job later in life
- Science-related jobs pay higher wages
- Science-related jobs require long years of study
- Science-related jobs are monotonous
- It is difficult have a fulfilling family life when working as a scientist
- Science-related jobs are rather solitary

C2. Among the following statements, which seem true to you, and which seem false?
[Possible answers: True/Somewhat true/Somewhat false/False]
- There are more men that women in science-related jobs
- Men are more gifted than women in mathematics
- Women and men are born with different brains

- Women don't really like science
- Women face discrimination in science-related jobs

C3. If you think of a [female/male] scientist, what adjectives comme to mind? Chose the adjective in each row that best fits your perception of a [female/male] scientist. (Check only only one box per row)
[Note: the female/male descriptor was randomly assigned across respondents]

☐ Interesting	OR	☐ Boring
☐ Repetitive	OR	☐ Creative
☐ Solitary	OR	☐ Sociable
☐ Stylish	OR	☐ Unfashionable
☐ Respected	OR	☐ Not very respected
☐ Shy	OR	☐ Extroverted
☐ Exemplary	OR	☐ Ordinary

Part D: Questions for Visited Classes [Treatment Group only]

D1. Did a female scientist from the L'Oreal program "For Girls in Science" visit your class?
☐ Yes ☐ No

D2. Did you enjoy this intervention?
☐ Yes ☐ No

D3. Would you say that this visit changed...
[Possible answers: Strongly agree/Somewhat agree/Disagree/Strongly disagree]
- your perception of science-related jobs?
- your interest in science-related jobs?
- your perception of women's place in science-related jobs?

D4. Would you say that this visit...
[Possible answers: Strongly agree/Somewhat agree/Disagree/Strongly disagree]
- gave you new ideas for your future?
- influenced your aspirations and choices of study?
- confirmed a choice you had already made?
- made you want to pursue science-related studies?

D5. Did you talk about this visit...
- with other students in your class?
  ☐ Yes ☐ No
- with students from other classes in your high school?
  ☐ Yes ☐ No
- with friends outside of your school?
  ☐ Yes ☐ No

D6. Have you been exposed to other science outreach programs?
- During this school year
  ☐ Yes ☐ No
- In the past
  ☐ Yes ☐ No ☐ I don't remember

Part D: External Environment [Control Group only]

D1. Did you talk about your choice of studies...

- with friends or other students in your class?
  ☐ Yes, a lot ☐ Yes, a little ☐ No
- with friends or other students in your high school who are not in your class?
  ☐ Yes, a lot ☐ Yes, a little ☐ No
- with friends or students outside of your school?
  ☐ Yes, a lot ☐ Yes, a little ☐ No

D2. Did you participate in science outreach programs and programs about science-related jobs via your high school (e.g., a visit from a scientist in your class, a science workshop, a visit to a museum or a laboratory)?

- I participated in such a program during this school year
  ☐ Yes ☐ No
- I will participate in such a program before the end of the school year
  ☐ Yes ☐ No
- I have participated in such a program in the past
  ☐ Yes ☐ No ☐ I don't remember

If so, what type of program was it (or will it be)? (check all boxes that apply)

☐ Science Fair (*Fête de la science*) workshop
☐ Scientist's visit to your class
☐ Visit to the Cité des Sciences or to the Palais de la Découverte
☐ Visit to a laboratory or to a company where scientists work
☐ Meeting with an association that promotes science
☐ Other

D3. Have you ever heard of this type of awareness programs? (check all boxes that apply)

☐ Yes, from other students in my high school
☐ Yes, from friends outside of my school
☐ Yes, from teachers
☐ Yes, from other people
☐ No, I haven't heard of such programs

D4. Do you know if other students in your high school received a classroom visit from a female or male scientist this year?

☐ No, I did not hear about it
☐ Yes, I vaguely heard about it
☐ Yes, I definitely heard about it

The questionnaire is now complete. Please check that you have answered ALL of the questions before handing the questionnaire to your teacher (in the sealed envelope). Thank you very much for your participation.

E Student-Level Administrative Data

This appendix describes the administrative data that we use to complement the information from the student survey (Section E.1) and provides details about the classification of STEM undergraduate programs (Section E.2).

E.1 Data Sources

For the purpose of the empirical analysis, we matched the data from our post-intervention student survey with three administrative datasets. These data were linked using an encrypted version of the French national student identifier (*Identifiant National Élève*).

High school enrollment data. Students' socio-demographic characteristics and enrollment status are obtained from the *Bases Élèves Académiques* (BEA) for academic years 2012/13 to 2016/17. These comprehensive administrative registers, which were provided by the three education districts of the Paris region (Paris, Créteil, and Versailles), cover the universe of students enrolled in the public and private high schools operating in the three districts. It also covers students enrolled in selective undergraduate programs, i.e., *Classes préparatoires aux Grandes Écoles* (CPGE) and *Sections de technicien supérieur* (STS), as these programs are located in high schools. The BEA data provide basic information on students' demographics (gender, date and country of birth, number of siblings), their parents' occupation, and detailed information on their enrollment status (school and class attended, elective courses taken). Students' socioeconomic status (SES) is measured using the French Ministry of Education's official classification, which uses the occupation of the child's legal guardian to define four groups of SES: high (company managers, executives, liberal professions, engineers, intellectual occupations, arts professions), medium-high (technicians and associate professionals), medium-low (farmers, craft and trades workers, service and sales workers), and low (manual workers and persons without employment).

University enrollment data. To track Grade 12 (science track) students' enrollment outcomes in non-selective undergraduate programs (*Licence*), we use a separate administrative data source, the *Système d'Information sur le Suivi de l'Étudiant* (SISE), which is managed by the Statistical Office of the French Ministry of Higher Education (Sous-Direction des Systèmes d'Information et des Études Statistiques, MESRI-SIES). This dataset, which covers the academic years 2012/13 to 2016/17, records all students enrolled in the French higher education system outside of CPGE and STS, except for the small fraction of students enrolled in undergraduate programs leading to paramedical and social care qualifications.

Data on student performance. The third dataset, the *Organisation des Concours et Examens Académiques et Nationaux* (OCEAN), contains students' individual exam results for the *Diplôme national du brevet* (DNB), which middle school students take at the end Grade 9, and for the *Baccalauréat*, which high school students take at the end of Grade 12. Access to this dataset, which covers the exams years 2010 to 2016, was provided by the Statistical Office of the French Ministry of Education (Direction de l'Évaluation, de la Prospective et de la Performance, MENJ-DEPP).

E.2 Classification of STEM Undergraduate Programs

The enrollment status of Grade 12 (science track)students in the year following the intervention, i.e., 2016/17, is measured by combing the information from the BEA and SISE datasets. For the

purpose of our analysis, we use two alternative classifications of STEM undergraduate programs, based on whether they are (i) selective or non-selective, and (ii) male- or female-dominated.

Selective vs. non-selective STEM programs.

- *Selective STEM*: This category includes all CPGE programs with a specialization in STEM, i.e., mathematics, physics and engineering science (MPSI), physics, chemistry and engineering science (PCSI), biology, chemistry, physics and earth sciences (BCPST), and physics, technology, and engineering science (PTSI). It also includes a small number of selective programs in engineering schools that recruit their students directly after high school graduation, as well as selective technical/vocational undergraduate programs (STS) that specialize in STEM fields.
- *Non-selective STEM*: This category includes non-selective university bachelor's degree programs (*Licence*) that specialize in STEM fields: math, physics, chemistry, earth and life sciences, and computer science. Undergraduate programs in medicine and pharmacy are not included in this category.

Male- vs. female-dominated STEM programs.

- *Male-dominated STEM*: We consider as male-dominated STEM programs those in which the share of female students is less than 50 percent. This category includes the selective programs (CPGE and STS) and non-selective programs (University) that specialize in mathematics, physics, chemistry, computer science, and engineering.
- *Female-dominated STEM*: This category includes both selective (CPGE and STS) and non-selective programs (*Licence*) that specialize in earth and life sciences.

If a student is enrolled in multiple higher education programs, we only consider the most selective among these programs, with CPGE taking precedence over STS, and STS taking precedence over university undergraduate degree programs.

F Summary Statistics and Balancing Tests

Table F3 – Experimental Sample: Summary Statistics

	High schools operating in the Paris region	Participating high schools	
		Classes selected for random assignment	Classes not selected for random assignment
	(1)	(2)	(3)
Number of high schools	489	98	96
Share private	0.22	0.17	0.08
Panel A. Grade 10			
Number of students	115,720	13,700	19,147
Number of classes	3,627	416	592
Female	0.525	0.529	0.525
Non-French	0.063	0.060	0.068
Age	15.14	15.13	15.14
High SES	0.403	0.381	0.361
Medium-high SES	0.118	0.128	0.127
Medium-low SES	0.239	0.241	0.248
Low SES	0.240	0.249	0.265
Number of siblings	1.44	1.49	1.50
Class size	32.22	33.25	32.48
DNB percentile rank in math	57.69	58.48	55.10
DNB percentile rank in French	57.23	57.85	55.75
Panel B. Grade 12 (science track)			
Number of students	38,582	5,751	5,623
Number of classes	1,267	185	179
Female	0.459	0.492	0.417
Age	17.11	17.12	17.10
Non-French	0.045	0.051	0.037
High SES	0.527	0.464	0.535
Medium-high SES	0.115	0.136	0.126
Medium-low SES	0.198	0.209	0.180
Low SES	0.160	0.192	0.160
Number of siblings	1.43	1.50	1.44
Class size	31.43	31.97	32.08
DNB percentile rank in math	76.25	74.06	76.20
DNB percentile rank in French	70.78	69.61	69.78

Notes: This table compares the characteristics of high schools that participated in the program evaluation to the characteristics of all general-track high schools operating in the Paris region. Among participating schools, Grade 10 and Grade 12 (science track) classes that were selected by principals for random assignment to treatment are compared to classes that were not selected. The summary statistics are computed from the *Bases Élèves académiques* of the three education districts of Paris, Créteil, and Versailles for the academic year 2015/16. French and math scores are from the exams of the *Diplôme national du brevet* (DNB) that middle school students take at the end of Grade 9.

Table F4 – Post-Intervention Role Model Survey: Summary Statistics

	Role model background				
	All	Profes-sionals	Resear-chers	Difference (3)−(2)	p-value of diff.
	(1)	(2)	(3)	(4)	(5)
A. Adults present during the intervention					
Teacher was present	0.890	0.883	0.896	0.014	0.773
Teacher's subject: science[a]	0.589	0.589	0.589	0.000	0.997
Teacher's gender: female	0.558	0.542	0.570	0.028	0.684
Teacher showed interest	0.692	0.635	0.736	0.102	0.115
Other adult present beside teacher	0.348	0.392	0.315	−0.077	0.236
B. General atmosphere during the intervention					
Students were very interested	0.423	0.425	0.422	−0.004	0.963
Students were very engaged in the discussion	0.386	0.378	0.392	0.014	0.838
Students were inattentive	0.134	0.165	0.110	−0.055	0.259
Powerpoint worked well	0.963	0.938	0.982	0.045	0.172
Videos worked well	0.888	0.891	0.886	−0.004	0.940
Logistical problems	0.160	0.185	0.140	−0.044	0.487
Talk interrupted due to discipline problems	0.068	0.079	0.060	−0.018	0.652
C. Students' responsiveness to topics addressed during the intervention					
Very responsive to "science is everywhere"	0.430	0.378	0.470	0.092	0.360
Very responsive to "jobs in science are fulfilling"	0.352	0.402	0.313	−0.088	0.333
Very responsive to "jobs in science are for girls too"	0.375	0.354	0.392	0.037	0.674
Very responsive to "jobs in science pay well"	0.387	0.263	0.476	0.213	0.042
Very responsive to the short videos	0.546	0.488	0.590	0.102	0.339
D. Overall impression of the role model					
Were gender stereotypes strong among students?					
Yes, very much	0.089	0.039	0.128	0.089	0.057
Rather yes	0.313	0.276	0.341	0.066	0.337
Rather no/not at all	0.598	0.685	0.530	−0.155	0.074
How did the classroom intervention go?					
Very well	0.556	0.535	0.572	0.037	0.670
Well	0.369	0.386	0.355	−0.030	0.716
Average/not so well/not well at all	0.075	0.079	0.072	−0.006	0.821
Was the intervention well suited to the students?					
Yes, very much	0.474	0.449	0.494	0.045	0.661
Rather yes	0.471	0.504	0.446	−0.058	0.574
Rather no/not at all	0.055	0.047	0.060	0.013	0.592
Number of role models	56	21	35		
Number of interventions	290	124	166		

Notes: The summary statistics are computed from the post-intervention role model survey that was administered online to collect feedback about the classroom visits. The unit of observation is a classroom intervention. [a] The science subjects taught in high school include mathematics, physics and chemistry, and earth and life sciences.

Table F5 – Treatment-Control Balance: Female Students

	Control group (1)	Treatment group (2)	Within school: Difference T−C (3)	Within school: p-value of diff. (4)
Panel A. Grade 10				
Student characteristics				
Age (years)	15.68	15.66	−0.01	0.369
Non-French	0.061	0.060	0.002	0.683
High SES	0.389	0.397	0.005	0.608
Medium-high SES	0.131	0.123	−0.008	0.242
Medium-low SES	0.241	0.229	−0.013	0.158
Low SES	0.239	0.251	0.016	0.079
Number of siblings	1.513	1.534	0.030	0.392
Class size	32.86	32.99	0.04	0.637
At least one science elective course	0.487	0.485	−0.006	0.795
At least one standard elective course	0.558	0.525	−0.041	0.070
DNB percentile rank in math	62.36	61.94	−0.43	0.522
DNB percentile rank in French	65.18	65.70	0.54	0.421
Test of joint significance	F-stat: 0.659 (p-value: 0.777)			
Predicted track in Grade 11				
Grade 11: Science track	0.309	0.305	−0.004	0.487
Grade 11: Science - general track	0.288	0.284	−0.004	0.544
Grade 11: Science - technological track	0.021	0.021	0.000	0.644
N	3,641	3,600	7,241	
Panel B. Grade 12 (science track)				
Student characteristics				
Age (years)	17.12	17.09	−0.04	0.036
Non-French	0.065	0.052	−0.017	0.040
High SES	0.444	0.459	0.025	0.080
Medium-high SES	0.139	0.128	−0.011	0.249
Medium-low SES	0.216	0.210	−0.010	0.376
Low SES	0.201	0.204	−0.005	0.619
Number of siblings	1.562	1.525	−0.057	0.054
Class size	31.76	32.17	0.38	0.247
DNB percentile rank in math	73.71	72.65	−0.47	0.511
DNB percentile rank in French	73.75	74.02	0.71	0.298
Test of joint significance	F-stat: 0.888 (p-value: 0.537)			
Predicted undergraduate major				
Major: STEM	0.288	0.286	0.000	0.865
Major: selective STEM	0.112	0.110	0.001	0.746
Major: male-dominated STEM	0.164	0.163	0.001	0.693
N	1,424	1,403	2,827	

Notes: Each row corresponds to a different linear regression with the dependent variable listed on the left, separately for female students in Grade 10 (Panel A) and in Grade 12 (Panel B). Columns 1 and 2 show the average value for students in the control and treatment groups, respectively. Column 3 reports the coefficient from the regression of each variable on the treatment group indicator, with the p-value reported in column 4. The regression controls for school fixed effects to account for the fact that randomization was stratified by school, and standard errors are adjusted for clustering at the unit of randomization (class). The F-statistic is from a test of the joint significance of the coefficients in a regression of the treatment group indicator on all student characteristics. High school tracks (Panel A) and undergraduate majors (Panel B) are predicted for each student using the coefficients from a linear regression of the corresponding binary variable (e.g., enrollment in a STEM major) on all student characteristics listed in the table. This model is fitted separately by grade level on the sample of students in the control group.

Table F6 – Treatment-Control Balance: Male Students

	Control group (1)	Treatment group (2)	Within school: Difference T−C (3)	Within school: p-value of diff. (4)
Panel A. Grade 10				
Student characteristics				
Age (years)	15.78	15.76	−0.01	0.329
Non-French	0.054	0.055	0.001	0.937
High SES	0.411	0.428	0.008	0.477
Medium-high SES	0.135	0.134	−0.008	0.317
Medium-low SES	0.236	0.220	−0.010	0.243
Low SES	0.218	0.218	0.010	0.256
Number of siblings	1.470	1.437	−0.013	0.642
Class size	32.71	32.92	0.04	0.687
At least one science elective course	0.576	0.597	0.008	0.739
At least one standard elective course	0.525	0.513	−0.023	0.319
DNB percentile rank in math	63.91	63.90	−0.35	0.610
DNB percentile rank in French	56.59	56.90	−0.12	0.845
Test of joint significance	F-stat: 0.551 (p-value: 0.868)			
Predicted track in Grade 11				
Grade 11: Science track	0.442	0.443	−0.002	0.781
Grade 11: Science - general track	0.334	0.337	0.000	0.979
Grade 11: Science - technological track	0.108	0.106	−0.002	0.575
N	3,160	3,299	6,459	
Panel B. Grade 12 (science track)				
Student characteristics				
Age (years)	17.17	17.12	−0.05	0.003
Non-French	0.042	0.045	0.005	0.440
High SES	0.463	0.488	0.031	0.038
Medium-high SES	0.133	0.142	0.008	0.377
Medium-low SES	0.216	0.193	−0.020	0.037
Low SES	0.188	0.177	−0.019	0.121
Number of siblings	1.458	1.452	−0.018	0.550
Class size	31.74	32.21	0.43	0.148
DNB percentile rank in math	74.63	75.18	0.78	0.211
DNB percentile rank in French	64.86	65.98	1.23	0.071
Test of joint significance	F-stat: 0.585 (p-value: 0.808)			
Predicted undergraduate major				
Major: STEM	0.475	0.477	0.002	0.528
Major: selective STEM	0.238	0.245	0.008	0.040
Major: male-dominated STEM	0.383	0.384	0.002	0.585
N	1,429	1,495	2,924	

Notes: Each row corresponds to a different linear regression with the dependent variable listed on the left, separately for male students in Grade 10 (Panel A) and in Grade 12 (Panel B). Columns 1 and 2 show the average value for students in the control and treatment groups, respectively. Column 3 reports the coefficient from the regression of each variable on the treatment group indicator, with the p-value reported in column 4. The regression controls for school fixed effects to account for the fact that randomization was stratified by school, and standard errors are adjusted for clustering at the unit of randomization (class). The F-statistic is from a test of the joint significance of the coefficients in a regression of the treatment group indicator on all student characteristics. High school tracks (Panel A) and undergraduate majors (Panel B) are predicted for each student using the coefficients from a linear regression of the corresponding binary variable (e.g., enrollment in a STEM major) on all student characteristics listed in the table. This model is fitted separately by grade level on the sample of students in the control group.

Table F7 – Compliance with Random Assignment

	All classes (1)	Classes assigned to: Control group (2)	Classes assigned to: Treatment group (3)
Panel A. Grade 10			
Number of classes visited by a role model	199	2	197
Number of classes not visited by a role model	217	205	12
Number of students	13,700	6,801	6,899
Student-level compliance with random assignment	0.97	0.99	0.94
Panel B. Grade 12 (science track)			
Number of classes visited by a role model	91	2	89
Number of classes not visited by a role model	94	90	4
Number of students	5,751	2,853	2,898
Student-level compliance with random assignment	0.97	0.98	0.95

Notes: This table reports compliance with the random assignment of Grade 10 and Grade 12 (science track) classes to the treatment and control groups. Two-way non-compliance was due to either classes in the treatment not being visited by a role model or to classes in the control group being visited by a role model.

Table F8 – Student Post-Treatment Survey: Response Rates

			Within school	
	Control group (1)	Treatment group (2)	Difference T−C (3)	p-value of diff. (4)
Panel A. Grade 10				
Survey response rate	0.879	0.905	0.026	0.026
			(0.012)	
Number of students	6,801	6,899	13,700	
Panel B. Grade 12 (science track)				
Survey response rate	0.909	0.912	0.005	0.693
			(0.012)	
Number of students	2,853	2,898	5,751	

Notes: This table reports the student survey response rate for students in the Grade 10 and Grade 12 (science track) classes that participated in the program. The response rates are computed based on the list of all students who were recorded in the *Bases Élèves académiques* as being enrolled in the participating classes during the academic year 2015/16. Columns 1 and 2 show the response rate of students in the control and treatment groups, respectively. Column 3 reports the coefficient from the regression of survey participation on the treatment group indicator, with p-values reported in column 4. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (in parentheses) are adjusted for clustering at the unit of randomization (class).

Table F9 – Treatment-Control Balance: Survey Respondents

	Control group (1)	Treatment group (2)	Within school: Difference T−C (3)	Within school: p-value of diff. (4)
Panel A. Grade 10				
Student characteristics				
Female	0.528	0.512	−0.014	0.160
Age (years)	15.73	15.70	−0.01	0.248
Non-French	0.056	0.056	0.003	0.528
High SES	0.402	0.417	0.005	0.496
Medium-high SES	0.134	0.130	−0.006	0.248
Medium-low SES	0.237	0.224	−0.009	0.200
Low SES	0.227	0.229	0.010	0.158
Number of siblings	1.489	1.483	−0.001	0.954
Class size	32.77	32.91	0.02	0.837
At least one science elective course	0.536	0.543	0.009	0.693
At least one standard elective course	0.539	0.519	−0.032	0.132
DNB percentile rank in math	63.65	63.42	−0.18	0.760
DNB percentile rank in French	61.53	61.80	0.08	0.893
Test of joint significance	F-stat: 0.634 (p-value: 0.813)			
Predicted track in Grade 11				
Grade 11: Science track	0.367	0.370	0.002	0.773
Grade 11: Science - general track	0.310	0.312	0.001	0.826
Grade 11: Science - technological track	0.057	0.058	0.000	0.854
N	5,981	6,245	12,226	
Panel B. Grade 12 (science track)				
Student characteristics				
Female	0.504	0.489	−0.014	0.319
Age (years)	17.13	17.09	−0.05	0.001
Non-French	0.053	0.046	−0.008	0.129
High SES	0.446	0.481	0.038	0.001
Medium-high SES	0.138	0.138	0.000	0.979
Medium-low SES	0.219	0.196	−0.022	0.001
Low SES	0.197	0.184	−0.016	0.086
Number of siblings	1.502	1.487	−0.021	0.355
Class size	31.69	32.12	0.30	0.314
DNB percentile rank in math	74.52	74.00	−0.09	0.874
DNB percentile rank in French	69.59	70.00	0.68	0.248
Test of joint significance	F-stat: 1.218 (p-value: 0.282)			
Predicted undergraduate major				
Major: STEM	0.395	0.395	0.001	0.807
Major: selective STEM	0.181	0.184	0.005	0.189
Major: male-dominated STEM	0.283	0.284	0.002	0.561
N	2,594	2,642	5,236	

Notes: Each row corresponds to a different linear regression with the dependent variable listed on the left, separately for students in Grade 10 (Panel A) and in Grade 12 (Panel B). The sample is restricted to students who answered the post-intervention survey. Columns 1 and 2 show the average value for students in the control and treatment groups, respectively. Column 3 reports the coefficient from the regression of each variable on the treatment group indicator, with the p-value reported in column 4. The regression controls for school fixed effects to account for the fact that randomization was stratified by school, and standard errors are adjusted for clustering at the unit of randomization (class). The F-statistic is from a test of the joint significance of the coefficients in a regression of the treatment group indicator on all student characteristics. High school tracks (Panel A) and undergraduate majors (Panel B) are predicted for each student using the coefficients from a linear regression of the corresponding binary variable (e.g., enrollment in a STEM major) on all student characteristics listed in the table. This model is fitted separately by grade level on the sample of students in the control group.

Table F10 – Balancing Test: High Schools Visited by Professionals and Researchers

	High school visited by		Difference (2)−(1)	p-value of diff.
	Researcher (1)	Professional (2)	(3)	(4)
Panel A. Grade 10				
Student characteristics				
Female	0.525	0.531	0.007	0.623
Age (years)	15.12	15.13	0.01	0.598
Non-French	0.065	0.057	−0.008	0.185
High SES	0.345	0.410	0.064	0.002
Medium-high SES	0.132	0.125	−0.007	0.322
Medium-low SES	0.250	0.235	−0.015	0.124
Low SES	0.272	0.231	−0.042	0.013
Number of siblings	1.482	1.488	0.007	0.862
Class size	33.38	33.14	−0.25	0.343
At least one science elective course	0.416	0.376	−0.040	0.250
At least one standard elective course	0.772	0.738	−0.034	0.197
DNB percentile rank in math	57.80	59.02	1.22	0.380
DNB percentile rank in French	56.77	58.71	1.93	0.120
Test of joint significance	F-stat: 1.165 (p-value: 0.306)			
Predicted track in Grade 11				
Grade 11: Science track	0.464	0.471	0.007	0.568
Grade 11: Science - general track	0.381	0.393	0.012	0.387
Grade 11: Science - technological track	0.083	0.078	−0.005	0.156
N	6,059	7,641	13.700	
Panel B. Grade 12 (science track)				
Student characteristics				
Female	0.474	0.505	0.032	0.114
Age (years)	17.14	17.11	−0.03	0.323
Non-French	0.057	0.046	−0.010	0.272
High SES	0.437	0.484	0.046	0.169
Medium-high SES	0.146	0.128	−0.018	0.138
Medium-low SES	0.213	0.205	−0.009	0.544
Low SES	0.203	0.184	−0.019	0.428
Number of siblings	1.454	1.532	0.079	0.100
Class size	32.67	31.44	−1.22	0.026
DNB percentile rank in math	72.96	74.90	1.94	0.213
DNB percentile rank in French	68.00	70.83	2.83	0.057
Test of joint significance	F-stat: 0.414 (p-value: 0.939)			
Predicted undergraduate major				
Major: STEM	0.382	0.383	0.001	0.926
Major: selective STEM	0.173	0.179	0.006	0.472
Major: male-dominated STEM	0.273	0.276	0.003	0.709
N	2,492	3,259	5.751	

Notes: Each row corresponds to a different linear regression with the dependent variable listed on the left, separately for students in Grade 10 (Panel A) and in Grade 12 (Panel B). Columns 1 and 2 show the average value for students whose high school was visited by a role model with a professional or a research background, respectively. Column 3 reports the coefficient from the regression of each variable on the treatment group indicator, with the p-value reported in column 4. Standard errors are adjusted for clustering at the unit of randomization (class). The F-statistic is from a test of the joint significance of the coefficients in a regression of the treatment group indicator on all student characteristics. High school tracks and undergraduate majors are predicted for each student using the coefficients from a linear regression of the corresponding binary variable (e.g., enrollment in a STEM major) on all student characteristics listed in the table. This model is fitted separately by grade level on the sample of students in the control group.

G Effects of Role Model Interventions: Additional Results

G.1 Student Perceptions

Table G11 – Gender Differences in Aptitude for Mathematics

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
Equal gender aptitude for math (index)	0.115	0.109***	0.000	−0.134	0.148***	0.000
		(0.025)			(0.030)	
M and W are born with different brains	0.211	−0.050***	0.000	0.209	−0.048***	0.000
		(0.010)	[0.001]		(0.011)	[0.001]
Men are more gifted in math than women	0.186	−0.026**	0.015	0.299	−0.048***	0.001
		(0.011)	[0.016]		(0.014)	[0.001]
N		6,475			5,751	
Panel B. Grade 12 (science track)						
Equal gender aptitude for math (index)	0.158	0.095***	0.001	−0.161	0.132***	0.001
		(0.028)			(0.040)	
M and W are born with different brains	0.143	−0.023**	0.026	0.180	−0.038***	0.006
		(0.010)	[0.026]		(0.014)	[0.013]
Men are more gifted in math than women	0.163	−0.038***	0.002	0.266	−0.028*	0.072
		(0.012)	[0.005]		(0.015)	[0.073]
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' perceptions regarding the aptitude of men and women for mathematics, separately by grade level and gender. The sample is restricted to students who completed the post-intervention questionnaire. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G12 – Taste for Science Subjects

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
Taste for science subjects (index)	−0.169	−0.038	0.294	0.197	−0.019	0.533
		(0.036)			(0.031)	
Enjoys math (z-score)	−0.147	−0.002	0.961	0.186	−0.002	0.935
		(0.034)	[0.961]		(0.031)	[0.935]
Enjoys physics-chemistry (z-score)	−0.170	−0.040	0.289	0.223	−0.022	0.505
		(0.038)	[0.578]		(0.033)	[0.935]
Enjoys earth and life sciences (z-score)	−0.042	−0.058	0.137	0.086	−0.027	0.443
		(0.039)	[0.548]		(0.035)	[0.935]
Enjoys science in general	0.661	−0.011	0.444	0.790	0.003	0.804
		(0.015)	[0.593]		(0.012)	[0.935]
N		6,475			5,751	
Panel B. Grade 12 (science track)						
Taste for science subjects (index)	−0.002	0.016	0.632	0.002	0.000	0.998
		(0.034)			(0.039)	
Enjoys math (z-score)	−0.097	0.067*	0.089	0.100	0.075*	0.063
		(0.040)	[0.357]		(0.040)	[0.203]
Enjoys physics-chemistry (z-score)	−0.089	−0.001	0.984	0.102	−0.021	0.598
		(0.044)	[0.984]		(0.040)	[0.599]
Enjoys earth and life sciences (z-score)	0.203	−0.030	0.435	−0.215	−0.059	0.318
		(0.038)	[0.871]		(0.059)	[0.424]
Enjoys science in general	0.918	−0.001	0.887	0.930	0.013	0.101
		(0.009)	[0.984]		(0.008)	[0.203]
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' taste for science subjects taught at school, separately by grade level and gender. The sample is restricted to students who completed the post-intervention questionnaire. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G13 – Math Self-Concept

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
Math self-concept (index)	−0.198	−0.008	0.806	0.231	0.039	0.217
		(0.031)			(0.032)	
Self-assessed math performance (z-score)	−0.127	−0.016	0.634	0.168	0.021	0.502
		(0.034)	[0.634]		(0.032)	[0.642]
Lost in front of a math problem	0.553	0.010	0.478	0.344	−0.007	0.610
		(0.014)	[0.634]		(0.013)	[0.642]
Worried when thinking about math	0.617	−0.025*	0.052	0.420	−0.032**	0.028
		(0.013)	[0.109]		(0.015)	[0.111]
Can succeed in science subjects if puts in effort	0.843	0.018*	0.054	0.883	−0.004	0.642
		(0.009)	[0.109]		(0.008)	[0.642]
N		6,475			5,751	
Panel B. Grade 12 (science track)						
Math self-concept (index)	−0.184	0.050	0.202	0.187	0.072**	0.041
		(0.039)			(0.035)	
Self-assessed math performance (z-score)	−0.126	0.039	0.304	0.123	0.079**	0.038
		(0.038)	[0.406]		(0.038)	[0.077]
Lost in front of a math problem	0.486	−0.028	0.168	0.325	−0.028*	0.072
		(0.020)	[0.336]		(0.016)	[0.096]
Worried when thinking about math	0.560	−0.037**	0.048	0.384	−0.051***	0.002
		(0.019)	[0.193]		(0.016)	[0.007]
Can succeed in science subjects if puts in effort	0.942	−0.005	0.512	0.949	0.006	0.384
		(0.007)	[0.512]		(0.007)	[0.385]
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' math self-concept, separately by grade level and gender. The sample is restricted to students who completed the post-intervention questionnaire. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G14 – Science-Related Career Aspirations

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
Science-related career aspirations (index)	−0.103	0.012	0.695	0.120	0.007	0.801
		(0.030)			(0.029)	
Some jobs in science are interesting	0.845	0.019**	0.050	0.854	0.000	1.000
		(0.009)	[0.200]		(0.010)	[1.000]
Would consider a job in science	0.466	−0.004	0.776	0.587	0.023*	0.089
		(0.015)	[0.776]		(0.014)	[0.358]
Interested in at least one STEM job[a]	0.642	0.005	0.696	0.849	0.013	0.181
		(0.013)	[0.776]		(0.010)	[0.363]
Wages important in career choice (z-score)	−0.045	−0.012	0.682	0.038	0.007	0.792
		(0.029)	[0.776]		(0.027)	[1.000]
N		6,475			5,751	
Panel B. Grade 12 (science track)						
Science-related career aspirations (index)	−0.045	0.113***	0.002	0.046	0.050	0.131
		(0.037)			(0.033)	
Some jobs in science are interesting	0.961	0.013**	0.013	0.940	0.021***	0.005
		(0.005)	[0.026]		(0.008)	[0.022]
Would consider a job in science	0.721	0.031**	0.019	0.762	0.030**	0.029
		(0.013)	[0.026]		(0.014)	[0.058]
Interested in at least one STEM job[a]	0.817	0.000	0.964	0.899	−0.001	0.946
		(0.011)	[0.964]		(0.009)	[0.947]
Wages important in career choice (z-score)	−0.043	0.119***	0.002	0.037	0.049	0.111
		(0.038)	[0.007]		(0.031)	[0.149]
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' self-reported science-related career aspirations, separately by grade level and gender. The sample is restricted to students who completed the post-intervention questionnaire. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. [a]: The STEM occupations in the list were chemist, computer scientist, engineer, industrial designer, renewable energy technician, and researcher in biology. The non-STEM occupations were lawyer, pharmacist, physician, and psychologist.

G.2 Educational Choices

Table G15 – Grade 10 Students: Enrollment Status the Following Year (Detailed)

	Grade 10 students					
	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. General Track						
Grade 11: Science track	0.343	−0.004	0.701	0.436	−0.001	0.899
		(0.014)	[0.889]		(0.014)	[0.928]
Grade 11: Humanities track	0.127	−0.003	0.799	0.029	0.005	0.319
		(0.010)	[0.889]		(0.005)	[0.478]
Grade 11: Social sciences track	0.264	0.010	0.395	0.171	0.012	0.239
		(0.012)	[0.889]		(0.010)	[0.478]
Panel B. Technological Track						
Grade 11: STEM-oriented technological tracks (STI2D, STL)	0.027	−0.006	0.112	0.141	−0.010	0.235
		(0.004)	[0.672]		(0.009)	[0.478]
Grade 11: non-STEM technological tracks	0.179	−0.005	0.656	0.139	0.000	0.979
		(0.011)	[0.889]		(0.009)	[0.980]
Repeater or dropout	0.057	0.004	0.580	0.078	−0.010	0.208
		(0.008)	[0.889]		(0.008)	[0.478]
N		7,241			6,459	

Notes: This table reports estimates of the treatment effects of the role model interventions on Grade 10 students' enrollment outcomes in the academic year following the classroom interventions, i.e. 2016/17, separately by gender. The enrollment outcomes are measured using student-level administrative data. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G16 – Grade 12 Students: Enrollment Status the Following Year (Detailed)

	Grade 12 (science track) students					
	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. STEM undergraduate programs						
Selective programs						
Math, physics, engineering, computer science	0.090	0.028***	0.006	0.222	0.018	0.239
		(0.010)	[0.068]		(0.016)	[0.814]
Earth & life sciences	0.020	0.008	0.137	0.010	0.002	0.617
		(0.005)	[0.496]		(0.003)	[0.834]
Non-selective programs						
Math, physics, computer science	0.077	0.010	0.209	0.157	−0.002	0.884
		(0.008)	[0.496]		(0.012)	[0.885]
Earth and life sciences	0.103	−0.022**	0.014	0.081	−0.015*	0.053
		(0.009)	[0.075]		(0.008)	[0.585]
Panel B. Non-STEM undergraduate programs						
Selective programs						
Business and economics	0.021	0.003	0.566	0.017	0.005	0.219
		(0.004)	[0.692]		(0.004)	[0.814]
Humanities	0.014	−0.004	0.225	0.003	−0.001	0.470
		(0.003)	[0.496]		(0.001)	[0.834]
Non-selective programs						
Medicine and pharmacy	0.259	−0.005	0.722	0.108	0.006	0.573
		(0.015)	[0.795]		(0.011)	[0.834]
Law and economics	0.107	−0.008	0.478	0.079	0.002	0.758
		(0.011)	[0.658]		(0.008)	[0.834]
Humanities and psychology	0.080	−0.008	0.339	0.040	−0.006	0.296
		(0.009)	[0.623]		(0.006)	[0.814]
Sports studies	0.023	−0.004	0.460	0.052	−0.005	0.555
		(0.006)	[0.658]		(0.009)	[0.834]
Not enrolled in an undergraduate program	0.218	0.000	1.000	0.237	0.005	0.739
		(0.015)	[1.000]		(0.016)	[0.834]
N		2,827			2,924	

Notes: This table reports estimates of the treatment effects of the role model interventions on science track Grade 12 (science track) students' enrollment outcomes in the academic year following the classroom interventions, i.e. 2016/17, separately by gender. The enrollment outcomes are measured using student-level administrative data. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

G.3 Academic Performance

Table G17 – Grade 12 Students: Performance in *Baccalauréat* Exams

	Grade 12 (science track) students					
	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Baccalauréat percentile rank in math	46.21	0.693	0.469	47.47	1.661	0.105
		(0.957)	[0.626]		(1.024)	[0.210]
Baccalauréat percentile rank in French	54.37	−0.051	0.964	43.51	−0.331	0.680
		(1.113)	[0.964]		(0.803)	[0.680]
Baccalauréat percentile rank	53.52	−1.121	0.293	47.29	1.712*	0.100
		(1.066)	[0.626]		(1.040)	[0.210]
Obtained the Baccalauréat	0.928	−0.010	0.334	0.877	−0.005	0.623
		(0.010)	[0.626]		(0.010)	[0.680]
N		2,827			2,924	

Notes: This table reports estimates of the treatment effects of the role model interventions on Grade 12 (science track) students' performance on the *Baccalauréat* exams, separately by gender. The enrollment outcomes are measured using student-level administrative data. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

H Robustness Checks

Table H18 – Treatment Effects on Student Perceptions: Controlling for Baseline Characteristics

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
Positive perceptions of science-related careers (index)	−0.020	0.245*** (0.027)	0.000	0.023	0.162*** (0.027)	0.000
More men in science-related jobs	0.628	0.154*** (0.013)	0.000 [0.001]	0.629	0.170*** (0.014)	0.000 [0.001]
Equal gender aptitude for math (index)	0.115	0.111*** (0.024)	0.000 [0.001]	−0.134	0.142*** (0.030)	0.000 [0.001]
Women like science less than men	0.157	0.056*** (0.011)	0.000 [0.001]	0.198	0.101*** (0.013)	0.000 [0.001]
W face discrimination in science-related jobs	0.603	0.126*** (0.013)	0.000 [0.001]	0.527	0.154*** (0.014)	0.000 [0.001]
Taste for science subjects (index)	−0.169	−0.033 (0.031)	0.275	0.197	−0.021 (0.026)	0.431
Math self-concept (index)	−0.198	−0.001 (0.028)	0.981	0.231	0.033 (0.029)	0.250
Science-related careers aspirations (index)	−0.103	0.005 (0.029)	0.851	0.120	0.004 (0.027)	0.871
N		6,475			5,751	
Panel B. Grade 12 (science track)						
Positive perceptions of science-related careers (index)	−0.003	0.296*** (0.032)	0.000	0.003	0.171*** (0.033)	0.000
More men in science-related jobs	0.712	0.122*** (0.016)	0.000 [0.001]	0.717	0.149*** (0.015)	0.000 [0.001]
Equal gender aptitude for math (index)	0.158	0.078*** (0.028)	0.004 [0.005]	−0.161	0.124*** (0.042)	0.003 [0.004]
Women like science less than men	0.074	0.042*** (0.009)	0.000 [0.001]	0.146	0.073*** (0.015)	0.000 [0.001]
W face discrimination in science-related jobs	0.624	0.085*** (0.020)	0.000 [0.001]	0.600	0.074*** (0.018)	0.000 [0.001]
Taste for science subjects (index)	−0.002	0.018 (0.033)	0.583	0.002	0.014 (0.040)	0.733
Math self-concept (index)	−0.184	0.051 (0.035)	0.139	0.187	0.068** (0.033)	0.038
Science-related careers aspirations (index)	−0.045	0.106*** (0.037)	0.004	0.046	0.068* (0.035)	0.055
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' perceptions, separately by grade level and gender, and controlling for students' baseline characteristics. The sample is restricted to students who completed the post-intervention questionnaire. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. The regression further controls for the student characteristics listed in Table 1 in the main text. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table H19 – Effects on Students' Perceptions: Weighted by the Inverse Probability of Answering the Questionnaire

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
Positive perceptions of science-related careers (index)	−0.020	0.245***	0.000	0.023	0.163***	0.000
		(0.029)			(0.029)	
More men in science-related jobs	0.628	0.156***	0.000	0.629	0.170***	0.000
		(0.013)	[0.001]		(0.014)	[0.001]
Equal gender aptitude for math (index)	0.115	0.109***	0.000	−0.134	0.139***	0.000
		(0.025)	[0.001]		(0.030)	[0.001]
Women like science less than men	0.157	0.057***	0.000	0.198	0.103***	0.000
		(0.011)	[0.001]		(0.013)	[0.001]
W face discrimination in science-related jobs	0.603	0.127***	0.000	0.527	0.154***	0.000
		(0.013)	[0.001]		(0.014)	[0.001]
Taste for science subjects (index)	−0.169	−0.037	0.307	0.197	−0.021	0.513
		(0.037)			(0.032)	
Math self-concept (index)	−0.198	−0.011	0.733	0.231	0.030	0.360
		(0.031)			(0.033)	
Science-related career aspirations (index)	−0.103	0.009	0.767	0.120	0.004	0.897
		(0.030)			(0.029)	
N		6,475			5,751	
Panel B. Grade 12 (science track)						
Positive perceptions of science-related careers (index)	−0.003	0.300***	0.000	0.003	0.175***	0.000
		(0.034)			(0.033)	
More men in science-related jobs	0.712	0.125***	0.000	0.717	0.148***	0.000
		(0.016)	[0.001]		(0.016)	[0.001]
Equal gender aptitude for math (index)	0.158	0.084***	0.003	−0.161	0.140***	0.001
		(0.028)	[0.003]		(0.042)	[0.001]
Women like science less than men	0.074	0.041***	0.000	0.146	0.072***	0.000
		(0.009)	[0.001]		(0.015)	[0.001]
W face discrimination in science-related jobs	0.624	0.086***	0.000	0.600	0.078***	0.000
		(0.020)	[0.001]		(0.019)	[0.001]
Taste for science subjects (index)	−0.002	0.015	0.671	0.002	0.013	0.757
		(0.035)			(0.041)	
Math self-concept (index)	−0.184	0.051	0.193	0.187	0.081**	0.036
		(0.039)			(0.038)	
Science-related career aspirations (index)	−0.045	0.105***	0.003	0.046	0.059*	0.095
		(0.036)			(0.036)	
N		2,600			2,636	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' perceptions that account for survey non-response, separately by grade level and gender. The sample is restricted to students who completed the post-intervention questionnaire and observations are weighted by the inverse predicted probability of answering the questionnaire. Survey response is predicted for each student from a linear regression of the survey response indicator on all student characteristics listed in Table 1 in the main text as well as school fixed effects. This model is fitted separately by grade level, gender, and treatment assignment. Each row corresponds to a different weighted linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt and using the inverse predicted probability of survey response as regression weights. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table H20 – Treatment Effects on Enrollment Outcomes: Controlling for Baseline Characteristics

	Girls			Boys		
	Control group mean (1)	Treatment effect (LATE) (2)	p-value [q-value] (3)	Control group mean (4)	Treatment effect (LATE) (5)	p-value [q-value] (6)
Panel A. Grade 10						
All STEM tracks						
Grade 11: Science track	0.371	−0.006	0.583	0.578	−0.010	0.402
		(0.011)			(0.012)	
General vs. technological STEM track						
Grade 11: Science - general track	0.343	−0.001	0.909	0.436	0.000	0.984
		(0.011)	[0.909]		(0.011)	[0.984]
Grade 11: Science - technological track	0.027	−0.005	0.173	0.141	−0.010	0.230
		(0.004)	[0.347]		(0.008)	[0.461]
N		7,241			6,459	
Panel B. Grade 12 (science track)						
All undergraduate STEM majors						
Major: STEM	0.289	0.020	0.139	0.470	−0.002	0.925
		(0.014)			(0.019)	
Selective vs. non-selective STEM						
Major: selective STEM	0.110	0.031***	0.006	0.232	0.008	0.575
		(0.011)	[0.012]		(0.015)	[0.575]
Major: non-selective STEM	0.178	−0.011	0.333	0.239	−0.010	0.445
		(0.012)	[0.333]		(0.013)	[0.575]
Male- vs. female-dominated STEM						
Major: male-dominated STEM (math, physics, computer science)	0.166	0.034***	0.004	0.379	0.013	0.485
		(0.012)	[0.012]		(0.019)	[0.575]
Major: female-dominated STEM (earth and life sciences)	0.123	−0.015	0.169	0.091	−0.015	0.119
		(0.011)	[0.226]		(0.009)	[0.477]
N		2,827			2,924	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' enrollment outcomes in the academic year following the classroom interventions, i.e., 2016/17, separately by grade level and gender, and controlling for student baseline characteristics. The enrollment outcomes are measured using student-level administrative data. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. They are obtained from a regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. The regression further controls for the student characteristics listed in Table 1 in the main text. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report the cluster-robust p-value of the estimated treatment effect and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

I Randomization Inference

This appendix evaluates the robustness of our results to computing p-values using non-parametric randomization inference tests rather than model-based cluster-robust inference.

Method. Randomization inference, which was first proposed by Fisher (1935) and was later developed by Rosenbaum (2002), has been used in a number of recent RCT studies in economics and political science as an alternative to model-based inference (e.g., Bloom et al., 2006; Cohen and Dupas, 2010; Ichino and Schündeln, 2012; Fujiwara and Wantchekon, 2013). The advantage of this method is that it is valid for any sample size and can be used even when the number of randomization units is small. It should be stressed, however, that randomization inference has lower power than parametric approaches when the true effect is large because it puts not even minimal structure on the error term (see discussion in Bloom et al., 2006).

The intuition behind this approach is relatively straightforward. In RCTs, researchers know exactly how the randomization was performed. Randomization inference uses this knowledge to assess whether observed outcomes in a given sample are likely to have been observed by chance even if the treatment had no effect. This can be obtained numerically through Monte Carlo methods, by computing the treatment effects for varying random draws of the treatment assignment, whose data-generating process is known. This test is non-parametric since it does not make distributional assumptions.[A.3]

Implementation. The ITT effect under the observed assignment to treatment is estimated using the following reduced-form specification:

$$Y_{ics} = \alpha + \beta T_{cs} + \theta_s + \epsilon_{ics}, \tag{A.1}$$

where Y_{ics} denotes the observed outcome of student i in class c and high school s; T_{cs} denotes the observed treatment assignment of the student's class; and θ_s are school fixed effects. The ITT estimate under the observed treatment assignment is denoted by $\hat{\beta}$.

To conduct randomization inference, we proceed as follows. Taking into account the fact that randomization was stratified by school and grade level, we first re-assign treatment R =2,000 times among participating classes using the exact same stratified procedure.[A.4] Let $\{P^r\}_{r=1}^{R}$ denote the set of R random placebo assignments from the randomization process. We then re-estimate the ITT effects of these placebo treatments using the following reduced-form specification, which is estimated separately by grade level and gender:

$$Y_{ics} = \alpha_r + \beta_r P_{cs}^r + \lambda_s + \eta_{ics}, \qquad r = 1, ..., R, \tag{A.2}$$

where P_{cs}^r is a dummy variable indicating assignment to a placebo treatment group for the random draw r. School fixed effects, λ_s, are included to account for the fact that the randomization is stratified by school.

Since P_r is a randomly generated placebo, $\mathbb{E}(\beta_r) = 0$. Let $F(\hat{\beta}_r)$ denote the empirical c.d.f. of all elements of $\{P_r\}_{r=1}^{R}$. We test the null hypothesis that the program had no effect on outcome Y by checking if the ITT estimate that we obtain for the observed treatment assignment is in the tails of the distribution of placebo treatments. We can reject H_0: $\hat{\beta} = 0$ with a confidence level of $1 - \alpha$ if $\hat{\beta} \leq F^{-1}\left(\frac{\alpha}{2}\right)$ or $\hat{\beta} \geq F^{-1}\left(1 - \frac{\alpha}{2}\right)$. Since the placebo assignments only vary across randomization units (classes), this method accounts for correlation within units.

[A.3] For more details on randomization inference and permutation tests, see Rosenbaum (2010) and Imbens and Rubin (2015).

[A.4] See Paz and West (2019) for the number of draws to be used.

Following (Davison and Hinkley, 1997, chap. 4), we compute the p-values from a two-sided randomization inference test of zero treatment effects as follows:

$$p = \frac{1 + \sum_{r=1}^{R} \mathbb{1}(|\hat{\beta}_r| \geq |\beta|)}{1 + R}, \tag{A.3}$$

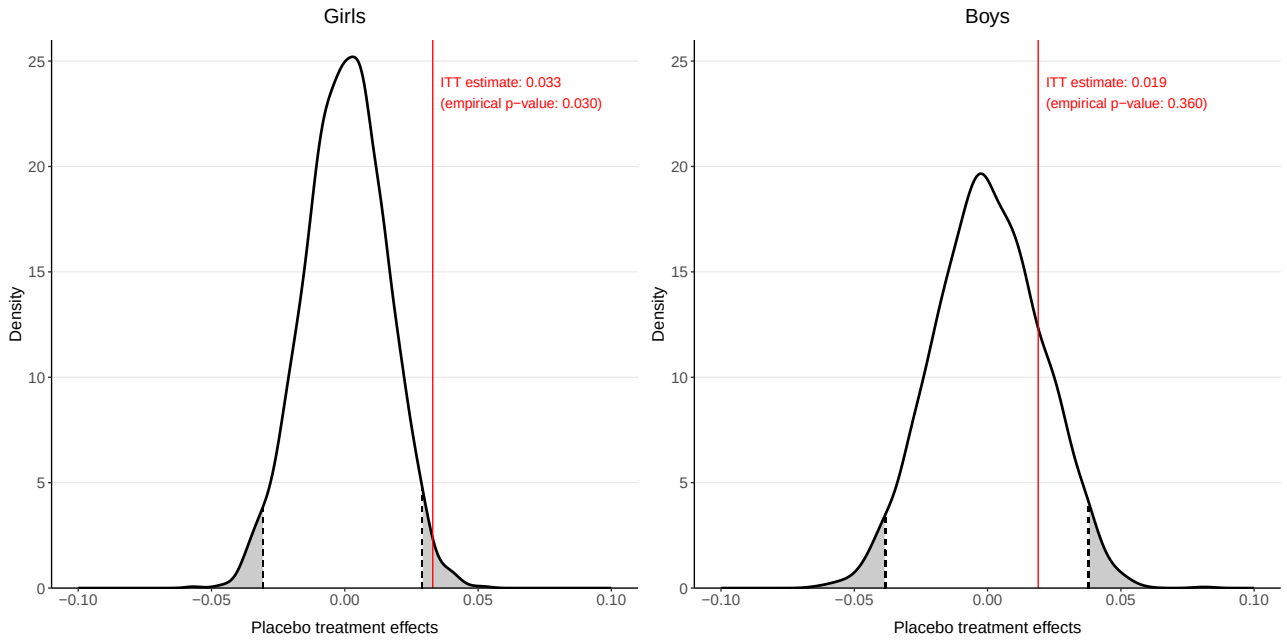
where $\mathbb{1}(\cdot)$ denotes the indicator function.

Results. Figure I5 illustrates the method by showing the empirical p.d.f. of the placebo ITT effects on selective and male-dominated STEM enrollment for girls in Grade 12 (science track), which are estimated using Equation (A.2). For each outcome, the vertical bar denotes the value of the ITT estimate for the observed assignment, which is obtained using Equation (A.1).

Table I21 presents the results of randomization inference tests of the hypotheses that the program had no effect on student perceptions and enrollment outcomes, separately by grade level and gender. The ITT estimates $\hat{\beta}$ are shown in columns 1 and 4. The associated cluster-robust model-based p-values are reported and columns 2 and 5, while the randomization inference p-values based on Equation (A.3) are in columns 3 and 6.

The results of the randomization inference tests yield p-values that are generally close to the cluster-robust model-based p-values. Although they tend to be slightly more conservative, they confirm the program's statistically significant effects on enrollment in selective and male-dominated STEM programs for girls in Grade 12.

(a) Enrollment in selective STEM programs



(b) Enrollment in male-dominated STEM programs

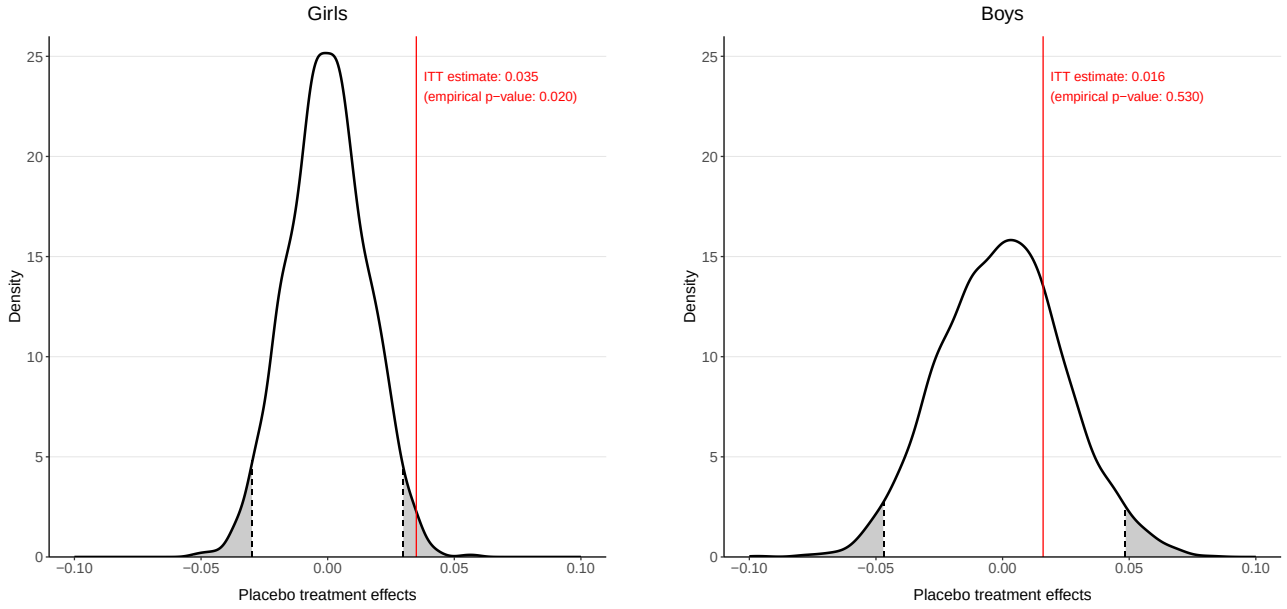


Figure I5 – Randomization Inference: Distribution of Placebo Treatment Effects on Enrollment in Selective and Male-Dominated STEM Undergraduate Programs, Grade 12

Notes: The figure shows the distribution of 2,000 placebo treatment effects (ITT) on the probability of enrolling in selective STEM (Panel a) and male-dominated STEM (Panel b) undergraduate programs for students in Grade 12 (science track), separately by gender. In each graph, the solid vertical line denotes the value of the ITT estimates for the observed assignment. The shaded area represents the two-sided (empirical) p-value of 0.05.

Table I21 – Randomization Inference for Intention-to-Treat estimates

	Girls			Boys		
	ITT	p-value: model-based	p-value: rand. inference	ITT	p-value: model-based	p-value: rand. inference
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Grade 10						
Student perceptions						
Positive perceptions of science-related careers (index)	0.226	0.000	0.000	0.156	0.000	0.000
More men in science-related jobs	0.145	0.000	0.000	0.157	0.000	0.000
Equal gender aptitude for math (index)	0.101	0.000	0.000	0.138	0.000	0.000
Women like science less than men	0.054	0.000	0.000	0.096	0.000	0.000
Women face discrimination in science-related careers	0.118	0.000	0.000	0.143	0.000	0.000
Taste for science subjects (index)	−0.035	0.298	0.340	−0.018	0.537	0.560
Math self-concept (index)	−0.007	0.808	0.820	0.037	0.221	0.280
Science-related career aspirations (index)	0.011	0.697	0.720	0.007	0.803	0.830
Enrollment outcomes						
Grade 11: Science track	−0.007	0.589	0.640	−0.006	0.679	0.710
Grade 11: Science - general track	−0.002	0.889	0.920	0.004	0.775	0.800
Grade 11: Science - technological track	−0.005	0.115	0.170	−0.010	0.240	0.300
N	7,241			6,459		
Panel B. Grade 12 (science track)						
Student perceptions						
Positive perceptions of science-related careers (index)	0.293	0.000	0.000	0.145	0.000	0.000
More men in science-related jobs	0.118	0.000	0.000	0.140	0.000	0.000
Equal gender aptitude for math (index)	0.090	0.001	0.020	0.124	0.002	0.020
Women like science less than men	0.042	0.000	0.000	0.069	0.000	0.000
Women face discrimination in science-related careers	0.090	0.000	0.000	0.068	0.000	0.000
Taste for science subjects (index)	0.015	0.640	0.740	0.000	0.998	1.000
Math self-concept (index)	0.047	0.214	0.360	0.068	0.044	0.140
Science-related career aspirations (index)	0.106	0.003	0.020	0.047	0.141	0.270
Enrollment outcomes						
Undergraduate major: STEM	0.022	0.091	0.220	0.003	0.889	0.920
Undergraduate major: selective STEM	0.033	0.002	0.030	0.019	0.208	0.360
Undergraduate major: non-selective STEM	−0.010	0.328	0.480	−0.016	0.220	0.370
Undergraduate major: male-dominated STEM	0.035	0.002	0.020	0.016	0.397	0.530
Undergraduate major: female-dominated STEM	−0.014	0.162	0.320	−0.013	0.128	0.270
N	2,827			2,924		

Notes: This table presents the results of randomization inference tests of the hypotheses that the program had no effect on student perceptions and enrollment outcomes. We randomly re-assigned treatment 2,000 times among participating classes within each school and grade level, and re-estimated the ITT effects of these placebo treatments. The ITT estimates under the observed assignment are reported in columns 1 and 4 separately by gender. The associated cluster-robust model-based p-values are shown in columns 2 and 5. The randomization inference p-values are reported in columns 3 and 6. Following Davison and Hinkley (1997), they are computed from a two-sided randomization inference test of zero treatment effects as $p = \left(1 + \sum_{r=1}^{R} \mathbb{1}(|\hat{\beta}_r| \geq |\beta|)\right)/(1+R)$, where $\{\hat{\beta}_r\}_{r=1}^{R}$ is the set of R placebo ITT estimates, $\hat{\beta}$ is the ITT estimate under the observed assignment, and $\mathbb{1}(\cdot)$ denotes the indicator function.

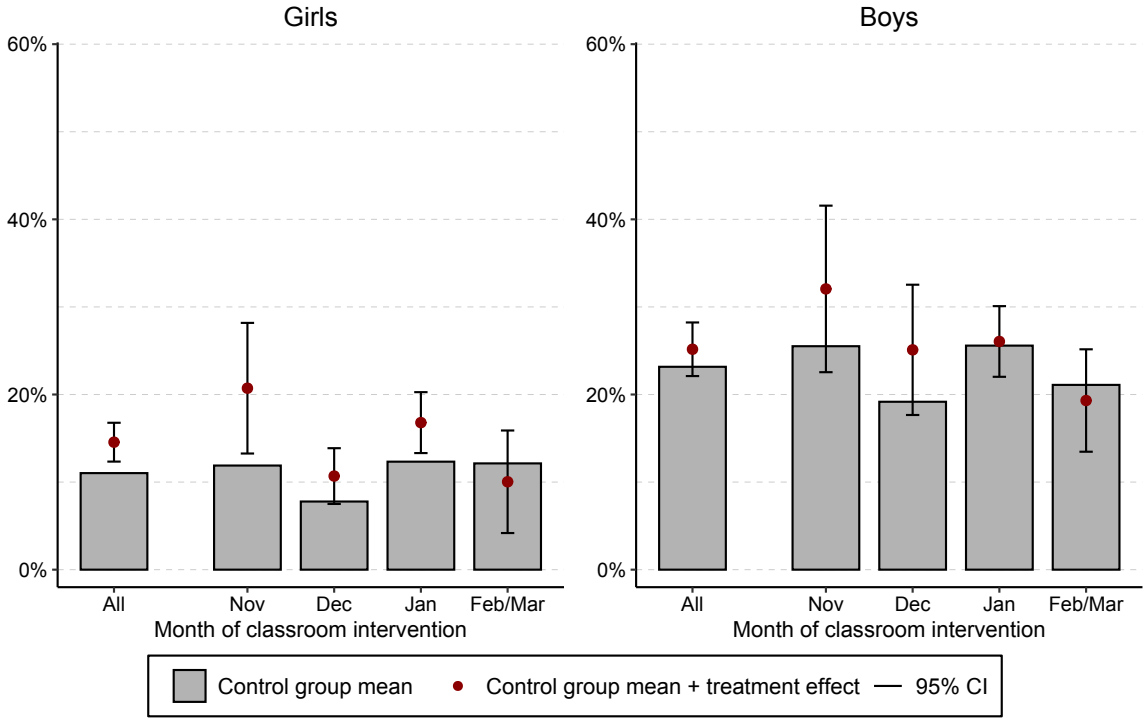
J Persistence of Effects and Timing of Visits

Table J22 – Persistence of Effects on Student Perceptions

	Girls			Boys		
	Months since intervention			Months since intervention		
	1 to 2 months (1)	3 to 4 months (2)	5 to 6 months (3)	1 to 2 months (4)	3 to 4 months (5)	5 to 6 months (6)
Panel A : Grade 10						
Positive perceptions of science-related careers (index)	0.413***	0.200***	0.143*	0.192***	0.168***	0.049
	(0.057)	(0.037)	(0.077)	(0.053)	(0.036)	(0.083)
More men in science-related jobs	0.170***	0.154***	0.164***	0.209***	0.163***	0.116***
	(0.021)	(0.017)	(0.033)	(0.022)	(0.018)	(0.039)
Equal gender aptitude for math (index)	0.179***	0.101***	0.019	0.244***	0.122***	0.090
	(0.047)	(0.032)	(0.065)	(0.053)	(0.040)	(0.069)
Women like science less than men	0.047**	0.067***	0.041	0.131***	0.107***	0.017
	(0.022)	(0.014)	(0.026)	(0.020)	(0.016)	(0.040)
W face discrimination in science-related careers	0.158***	0.135***	0.081**	0.162***	0.174***	0.110***
	(0.022)	(0.017)	(0.039)	(0.026)	(0.017)	(0.036)
Taste for science subjects (index)	0.088	−0.035	−0.053	0.043	−0.008	0.043
	(0.075)	(0.043)	(0.075)	(0.058)	(0.041)	(0.072)
Math self-concept (index)	−0.029	0.006	0.044	−0.041	0.103***	0.088
	(0.057)	(0.039)	(0.080)	(0.063)	(0.039)	(0.090)
Science-related career aspirations (index)	0.088	−0.002	0.010	0.000	0.022	0.010
	(0.057)	(0.036)	(0.062)	(0.051)	(0.038)	(0.072)
N	1,729	3,716	831	1,677	3,318	693
Panel B : Grade 12 (science track)						
Positive perceptions of science-related careers (index)	0.442***	0.253***	0.353***	0.182***	0.169***	0.003
	(0.053)	(0.043)	(0.118)	(0.061)	(0.044)	(0.095)
More men in science-related jobs	0.128***	0.107***	0.208***	0.114***	0.159***	0.208***
	(0.031)	(0.019)	(0.060)	(0.023)	(0.021)	(0.046)
Equal gender aptitude for math (index)	0.077	0.138***	0.020	0.218***	0.106**	0.044
	(0.067)	(0.033)	(0.094)	(0.081)	(0.051)	(0.123)
Women like science less than men	0.067***	0.040***	0.032*	0.042	0.077***	0.144***
	(0.021)	(0.011)	(0.018)	(0.029)	(0.019)	(0.032)
W face discrimination in science-related jobs	0.104***	0.102***	0.087	0.083***	0.085***	−0.011
	(0.038)	(0.023)	(0.072)	(0.027)	(0.024)	(0.062)
Taste for science subjects (index)	−0.063	−0.028	0.258***	0.030	0.010	−0.090
	(0.071)	(0.045)	(0.060)	(0.079)	(0.049)	(0.111)
Math self-concept (index)	0.043	0.001	0.169	−0.022	0.114**	0.126
	(0.065)	(0.053)	(0.122)	(0.054)	(0.046)	(0.149)
Science-related career aspirations (index)	−0.005	0.123***	0.231***	0.007	0.048	0.098
	(0.077)	(0.045)	(0.077)	(0.046)	(0.046)	(0.118)
N	689	1,468	394	717	1,514	370

Notes: This table reports estimates of the treatment effects of the role model interventions on student perceptions, separately by grade level, gender, and intervals of elapsed time between the classroom intervention and the student survey. The sample is restricted to students who completed the post-intervention questionnaire. Each coefficient is obtained from a linear regression of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class).

(a) Enrollment in selective STEM



(b) Enrollment in male-dominated STEM

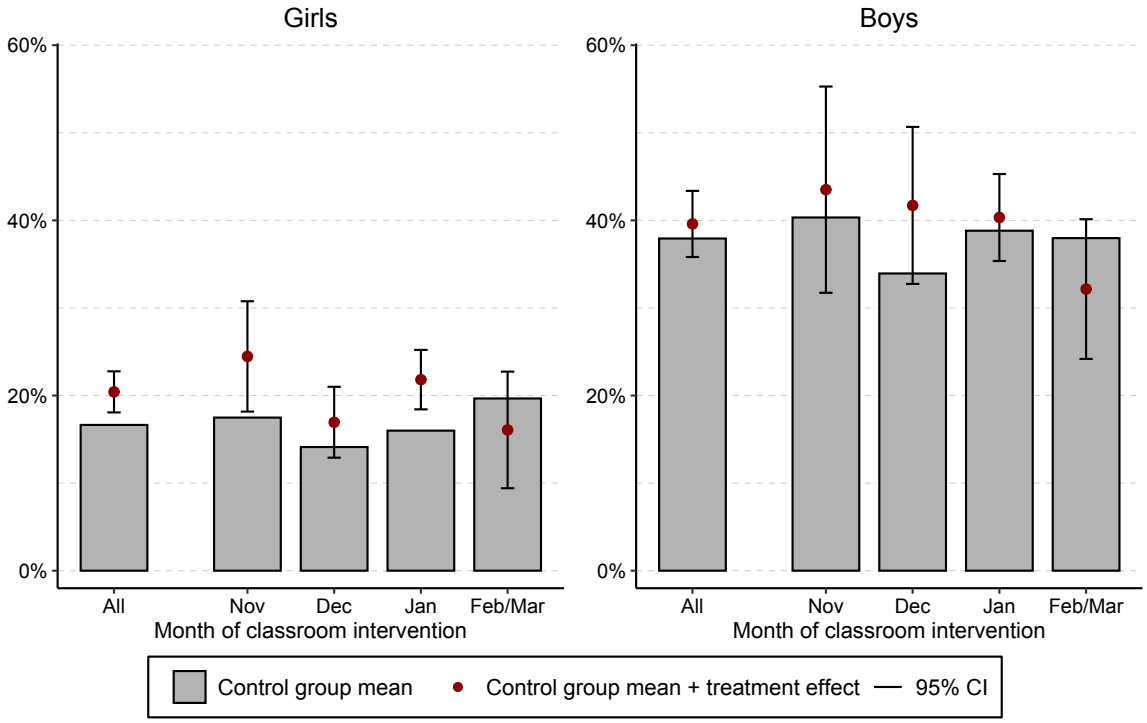


Figure J6 – Grade 12 Students: Enrollment in Selective and Male-Dominated STEM Undergraduate Programs, by Gender and Month of Classroom Intervention

Notes: The figure shows the fraction of Grade 12 (science track) students who enrolled in selective (Panel a) and in male-dominated (Panel b) STEM undergraduate programs after graduating for high school, separately for girls (left panel) and for boys (right panel). The filled bars indicate the baseline enrollment rates among students in the control group, both overall and separately by month of classroom intervention. The solid dots show the estimated treatment effects (added to the control group means) with 95 percent confidence intervals denoted by vertical capped bars. The treatment effects are estimated from separate regressions of the outcome of interest on a classroom visit indicator, using treatment assignment as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors are adjusted for clustering at the unit of randomization (class).

K Spillover Effects

This appendix investigates whether the program could have had spillover effects for students who were not exposed to the role model interventions in the schools participating in the evaluation. Section K.1 provides survey evidence suggesting that the scope for spillover effects was relatively limited. Section K.2 describes the difference-in-differences (DiD) approach that we use to estimate the magnitude of spillovers, the results of which point to non-statistically significant effects.

K.1 Survey Evidence

To get some sense of the scope for spillover effects in the context of our study, we included in the last section of the survey a series of questions asking students in the treatment group whether they had talked about the classroom interventions with their classmates, with schoolmates from other classes, or with friends from other schools (see the survey questionnaire in Appendix D). We also asked students in the control group whether they had heard about a science-related awareness-raising program and, more specifically, whether they knew about other classes in the school being visited by a female or male scientist.

Overall, the summary statistics from the survey data suggest relatively limited opportunities for spillover effects (see Appendix Table K23). In the treatment group, 58 percent of Grade 10 students and 63 percent of science track Grade 12 students report having talked about the classroom intervention with their classmates, but only 24 percent (27 percent) with schoolmates from other classes, and 20 percent with students from other schools. Interestingly, these proportions are higher for girls than for boys: in Grade 10, 66 percent of girls in the treatment group report having discussed the program with their classmates and 28 percent with schoolmates from other classes vs. respectively 50 percent and 20 percent among boys; in Grade 12, 70 percent of girls in the treatment group report having discussed the program with their classmates and 33 percent with schoolmates from other classes vs. respectively 56 percent and 21 percent among boys.

In the control group, only 14 percent of students in Grade 10 report having heard of classroom visits in other classes, mostly in a vague manner (12 percent). In Grade 12, students in the control group are more likely to report being at least vaguely aware of such visits (34 percent), but less than 5 percent of boys and girls have a precise recollection. Gender differences in these proportions are small and barely statistically significant. The fact that students in Grade 12 are more likely to report being aware of classroom visits could be at least partly due to the fact that the share of students assigned to the treatment group among all students from the same grade level was typically larger in Grade 12 than in Grade 10, on average 32 percent vs. 25 percent. Despite these differences, the overall picture that emerges from the survey is that students in the control group had only limited awareness of the classroom interventions in other classes.

K.2 Differences-in-Differences Estimates of Spillover Effects

We complement the survey evidence by investigating more formally whether the role model interventions could have affected the higher education choices of Grade 12 students whose classes were not assigned to the treatment group. These students are either in the classes that were not selected by school principals to participate in the program evaluation or in the participating classes that were randomly assigned to the control group.

Our experimental design does not include a "super control" group composed of students enrolled in schools randomly chosen to have zero probability of assignment to the treatment among the classes selected by school principals. Spillover effects cannot therefore be identified by comparing the control group classes in participating schools with such supercontrol group

classes, as in the design pioneered by Duflo and Saez (2003).[A.5] Instead, our approach builds on the following intuition: for schools that participated in the evaluation, the random assignment of treatment to participating classes makes it possible to estimate the average outcome that would have been observed if *all* students from these schools had only been exposed to the spillover effects of role model interventions, without being *directly* exposed to a female role model. This unobserved "spillover-only" counterfactual can be estimated at the school level using an appropriately weighted average of non-treated classes: it suffices to compute the weighted average outcome of students in the non-participating classes and in the participating classes that were randomly assigned to the control group, with respective weights equal to the share of participating and of non-participating classes in the school. Average spillover effects can then be estimated by comparing this "spillover-only" counterfactual to a "no-treatment" counterfactual. This second counterfactual is constructed under the assumption that absent treatment, mean outcomes in participating school would have followed the same evolution as in non-participating schools. Having verified that this common trends assumption is satisfied in the pre-treatment period 2012–2014, we implement a difference-in-differences estimator that identifies the difference between the "spillover-only" and the "no-treatment" counterfactuals. This approach, which is graphically illustrated in Figure K7, enables us to estimate the average spillover effects of role model interventions in the participating schools.

Notations. We are interested in measuring the spillover effects of classroom visits. We denote by D_s a binary indicator for a student's school s being visited by a female role model and by D_{cs} a binary indicator for a role model intervention taking place in the student's class c. We consider two time periods, represented by a binary indicator $T \in \{0, 1\}$, with classroom visits taking place in period 1 only. For a given realization of the treatment assignment (d_s, d_{cs}), the potential outcome for student i in school s, class c, and time t is denoted by $Y_{icst}(d_s, d_{cs})$.

We use the binary indicator G_s to indicate whether school s participated in the experiment and we denote the sets of participating and non-participating schools by $\mathcal{S}_1$ and $\mathcal{S}_0$, respectively. The number of participating (non-participating) schools is denoted by M_1 (M_0). Only a subset of the classes in participating schools were (non-randomly) selected by the principals to participate in the experiment in period 1. The participation status of class c in school s is denoted by the binary indicator G_{cs}. Among participating classes ($G_{cs} = 1$), the binary indicator R_{cs} indicates whether the class was randomly assigned to the treatment group ($R_{cs} = 1$) or to the control group ($R_{cs} = 0$). The experimental setting therefore implies that $D_s = G_s \times T$ and $D_{cs} = R_{cs} \times T$. A student's observed outcome can then be written

$$Y_{icst} = D_s \cdot D_{cs} \cdot Y_{icst}(1,1) + D_s \cdot (1 - D_{cs}) \cdot Y_{icst}(1,0) + (1 - D_s) \cdot Y_{icst}(0,0). \tag{A.4}$$

To simplify notation, we assume that each school has the same number of students, N, and that the number of students is the same in both periods.

Let $\overline{Y}_{s,t}(0,0)$ denote the average *potential* outcome of students in school s and year t under no treatment. This average potential outcome corresponds to the case in which no student from school s in year t is exposed to either the direct or spillover effects of classroom visits, i.e.,

$$\overline{Y}_{s,t}(0,0) = \frac{1}{N} \sum_{i=1}^{N} Y_{icst}(0,0). \tag{A.5}$$

[A.5] Vazquez-Bare (2018) develops a potential-outcome-based nonparametric framework to identify spillover effects in randomized experiments where units are clustered, without requiring a specific experimental design. This approach, however, cannot be easily adapted to our setting since it requires that the treatment is assigned at the individual level within clusters (schools), not at the group level (classes), in order to exploit variation in all the possible configurations of own and neighbors' observed treatment assignments.

Let $\overline{Y}_{s,t}(1,0)$ denote the average *potential* outcome of students in school s and year t in the (non-feasible) scenario in which all students in school s are only exposed to the spillover effects of role model interventions in other classes, without themselves being visited by a female role model. This "spillover-only" average potential outcome is defined as follows:

$$\overline{Y}_{s,t}(1,0) = \frac{1}{N}\sum_{i=1}^{N} Y_{icst}(1,0). \tag{A.6}$$

Our parameter of interest is the expected average spillover effect of classroom visits for the students in participating schools in period 1, i.e.,

$$\Delta = \mathbb{E}\left(\frac{1}{M_1}\sum_{s\in\mathcal{S}_1}\left(\overline{Y}_{s,1}(1,0) - \overline{Y}_{s,1}(0,0)\right)\right). \tag{A.7}$$

This parameter can be interpreted as the average effect for students in participating schools of being only exposed to the indirect effects of classroom visits compared to the counterfactual of no classroom visit in the school.

Identification of spillover effects. Let $\overline{Y}_{s,t}$ denote the mean *observed* outcome for students in school s and year t, i.e.,

$$\overline{Y}_{s,t} = \frac{1}{N}\sum_{i=1}^{N} Y_{icst}. \tag{A.8}$$

For non-participating schools in periods 0 and 1 and for participating schools in period 0, this mean observed outcome is in expectation equal to the expected average potential outcome under no treatment. Indeed, Equations (A.4), (A.5), and (A.8) imply that

$$\mathbb{E}(\overline{Y}_{s,t}) = \mathbb{E}\left(\overline{Y}_{s,t}(0,0)\right) \text{ if } s\in\mathcal{S}_0 \text{ and } t\in\{0,1\} \text{ or if } s\in\mathcal{S}_1 \text{ and } t=0 \tag{A.9}$$

For each school $s\in\mathcal{S}_1$ that participated in the evaluation, we consider the following partition of students in period 1: let $\mathcal{C}_s^0$, $\mathcal{C}_s^C$, and $\mathcal{C}_s^T$ denote respectively (i) the students in the classes that did not participate in the evaluation ($G_s=0$), (ii) the students in the participating classes that were randomly assigned to the control group ($G_s=1$ and $R_{cs}=0$), and (iii) the students in the participating classes that were randomly assigned to the treatment group ($G_s=1$ and $R_{cs}=1$). By definition, the number of students in each group, which we denote by N_s^0, N_s^C and N_s^T respectively, is such that $N = N_s^0 + N_s^C + N_s^T$.

For the purpose of estimating spillover effects, we construct a mean counterfactual outcome for participating schools in period 1, which we denote by $\widetilde{Y}_{s,1}$. As shown in Proposition 1 below, the expected value of $\widetilde{Y}_{s,1}$ coincides with the expected average potential outcome of students in school s and period 1 if all students had only been exposed to the spillover effects of classroom visits in other classes, without being themselves directly exposed to a female role model. This counterfactual outcome ignores classes in the treatment group and is defined as a weighted average of the observed outcomes of students in the non-participating classes and the control group classes (see Figure K7):

$$\widetilde{Y}_{s,1} = \frac{1}{N}\left(\sum_{i\in\mathcal{C}_s^0} Y_{ics1} + \left(1+\frac{N_s^T}{N_s^C}\right)\sum_{i\in\mathcal{C}_s^C} Y_{ics1}\right), \quad s\in\mathcal{S}_1. \tag{A.10}$$

The intuition is as follows. The "spillover only" counterfactual measured at the school level cannot be recovered from the non-participating classes only, since these classes were

not randomly selected by school principals. However, having noted that the mean observed outcome of students in the control group is an unbiased estimator of the mean (unobserved) "spillover-only" outcome for students in the treatment group, one can reconstruct the school-level "spillover-only" counterfactual by restricting the set of students to those in non-participating classes and control group classes. To estimate the mean outcome that would have been observed if all students had only been exposed to the spillover effects of classroom visits, it suffices to reweight students in the control group so that they match the total number of students in the participating classes (i.e., treatment and control), and then combine this reweighted sample with the sample of students in non-participating classes to compute the average outcome.

Assumption 1. *Random assignment of treatment to participating classes.*

$$\mathbb{E}\left(\frac{1}{N_s^T}\sum_{i\in\mathcal{C}_s^T} Y_{ics1}(1,0)\right) = \mathbb{E}\left(\frac{1}{N_s^C}\sum_{i\in\mathcal{C}_s^C} Y_{ics1}(1,0)\right), \quad s \in \mathcal{S}_1.$$

Assumption 1 states that students in the treatment and control group classes of participating schools have the same expected average potential outcome under the "spillover-only" treatment. Our experimental design ensures that this assumption is satisfied.

Proposition 1. *Under Assumption 1, the counterfactual $\widetilde{Y}_{s,1}$ is an unbiased estimator of the expected average potential outcome of students in participating school s and period 1 under the "spillover-only" treatment, $\overline{Y}_{s,1}(1,0)$:*

$$\mathbb{E}(\widetilde{Y}_{s,1}) = \mathbb{E}\left(\overline{Y}_{s,1}(1,0)\right), \quad s \in \mathcal{S}_1.$$

Proof. From the definition of the "spillover-only" counterfactual in Equation (A.10), we have

$$\begin{aligned}
\mathbb{E}(\widetilde{Y}_{s,1}) &= \mathbb{E}\left(\frac{1}{N}\left(\sum_{i\in\mathcal{C}_s^0} Y_{ics1} + \left(1+\frac{N_s^T}{N_s^C}\right)\sum_{i\in\mathcal{C}_s^C} Y_{ics1}\right)\right) \\
&= \frac{1}{N}\left(\sum_{i\in\mathcal{C}_s^0}\mathbb{E}(Y_{ics1}(1,0)) + \sum_{i\in\mathcal{C}_s^C}\mathbb{E}(Y_{ics1}(1,0)) + \frac{N_s^T}{N_s^C}\sum_{i\in\mathcal{C}_s^C}\mathbb{E}(Y_{ics1}(1,0))\right) \\
&= \frac{1}{N}\left(\sum_{i\in\mathcal{C}_s^0}\mathbb{E}(Y_{ics1}(1,0)) + \sum_{i\in\mathcal{C}_s^C}\mathbb{E}(Y_{ics1}(1,0)) + \sum_{i\in\mathcal{C}_s^T}\mathbb{E}(Y_{ics1}(1,0))\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(Y_{ics1}(1,0)) \\
&= \mathbb{E}\left(\overline{Y}_{s,1}(1,0)\right).
\end{aligned}$$

The second equality follows from Equation (A.4), the third equality follows from Assumption 1, while the last equality follows from Equation (A.6). The key intuition for this result is that by virtue of the random assignment of treatment to participating classes, the mean observed outcome of students assigned to the control group is an unbiased estimator of the mean unobserved "spillover-only" outcome of students assigned to the treatment group. □

Identifying spillover effects requires comparing the "spillover-only" counterfactual with the "no-treatment" counterfactual. To this end, we define the following difference-in-differences estimator, which we denote by $\hat{\Delta}$:

$$\hat{\Delta} = \frac{1}{M_1}\sum_{s\in\mathcal{S}_1}(\widetilde{Y}_{s,1} - \overline{Y}_{s,0}) - \frac{1}{M_0}\sum_{s\in\mathcal{S}_0}(\overline{Y}_{s,1} - \overline{Y}_{s,0}). \tag{A.11}$$

This estimator compares the evolution of the mean outcome of students in participating schools between period 0 and period 1 (using the "spillover-only" counterfactual for period 1) with the corresponding evolution in non-participating schools.

Assumption 2. *Common trends between participating and non-participating schools.*

$$\mathbb{E}\left(\frac{1}{M_1}\sum_{s\in\mathcal{S}_1}\left(\overline{Y}_{s,1}(0,0)-\overline{Y}_{s,0}(0,0)\right)\right)=\mathbb{E}\left(\frac{1}{M_0}\sum_{s\in\mathcal{S}_0}\left(\overline{Y}_{s,1}(0,0)-\overline{Y}_{s,0}(0,0)\right)\right).$$

Assumption 2 states that in the absence of role model visits to the school, average outcomes in participating and non-participating schools would have followed parallel trends. Although this assumption cannot be directly tested, it can be indirectly assessed by comparing the evolution of mean outcomes in participating and non-participating schools in the pre-intervention period.

Proposition 2. *Under Assumptions 1 and 2, $\hat{\Delta}$ is an unbiased estimator of the average spillover effect, Δ:*

$$\mathbb{E}(\hat{\Delta})=\Delta.$$

Proof. From the definition of the difference-in-differences estimator in Equation (A.11), we have

$$\begin{aligned}
\mathbb{E}(\hat{\Delta}) &= \mathbb{E}\left(\frac{1}{M_1}\sum_{s\in\mathcal{S}_1}\left(\widetilde{Y}_{s,1}-\overline{Y}_{s,0}\right)-\frac{1}{M_0}\sum_{s\in\mathcal{S}_0}\left(\overline{Y}_{s,1}-\overline{Y}_{s,0}\right)\right)\\
&= \mathbb{E}\left(\frac{1}{M_1}\sum_{s\in\mathcal{S}_1}\left(\overline{Y}_{s,1}(1,0)-\overline{Y}_{s,0}(0,0)\right)\right)-\mathbb{E}\left(\frac{1}{M_0}\sum_{s\in\mathcal{S}_0}\left(\overline{Y}_{s,1}(0,0)-\overline{Y}_{s,0}(0,0)\right)\right)\\
&= \mathbb{E}\left(\frac{1}{M_1}\sum_{s\in\mathcal{S}_1}\left(\overline{Y}_{s,1}(1,0)-\overline{Y}_{s,0}(0,0)\right)\right)-\mathbb{E}\left(\frac{1}{M_1}\sum_{s\in\mathcal{S}_1}\left(\overline{Y}_{s,1}(0,0)-\overline{Y}_{s,0}(0,0)\right)\right)\\
&= \mathbb{E}\left(\frac{1}{M_1}\sum_{s\in\mathcal{S}_1}\left(\overline{Y}_{s,1}(1,0)-\overline{Y}_{s,1}(0,0)\right)\right)\\
&= \Delta.
\end{aligned}$$

The second equality follows from Equation (A.9) and from Proposition 1, the third equality follows from Assumption 2 (common trends between participating and non-participating schools), while the last equality follows from Equation (A.7). □

Empirical specification. In the context of our study, the spillover effects estimator (A.11) can be conveniently implemented using a difference-in-differences regression specification. We apply this estimator to investigate whether the classroom interventions affected the college decisions of science track Grade 12 students whose classes were not visited by a female role model.

In our empirical application, we consider the four cohorts of Grade 12 students that were enrolled in the high schools of the Paris region in the year of the intervention (2015) and in the three preceding years (2012, 2013, and 2014).

One complication is that the "For Girls in Science" program was first implemented on a small scale in 2014, i.e., one year before the evaluation was conducted (in 2015). As a result, some of the schools that participated in the program evaluation, as well as some of the schools that did not participate in the evaluation, could have been visited by female role models in 2014. Although we cannot precisely identify these schools, the contamination effect is likely to be small since the interventions were carried out by a small number of role models and were not specifically

targeted at students enrolled in Grade 10 and Grade 12 (science track). Nonetheless, to ensure that our difference-in-differences estimates are not biased due to these prior interventions, we use 2012 as the reference year. The baseline differences between participating and non-participating schools are therefore measured at a point in time in which the program was not in place.

Let $\overline{Y}_{s,t}$ denote the average outcome of Grade 12 students in school s and year t. For each participating school $s \in \mathcal{S}_1$, we use Equation (A.10) to construct the "spillover-only" mean counterfactual outcome in 2015, which we denote by $\widetilde{Y}_{s,t}$. Our dependent variable, denoted by $\overline{Y}^*_{s,t}$, is then defined as follows:

$$\overline{Y}^*_{s,t} = \begin{cases} \widetilde{Y}_{s,t} & \text{if } s \in \mathcal{S}_1 \text{ and } t = 2015 \\ \overline{Y}_{s,t} & \text{otherwise} \end{cases}$$

The spillover effects of classroom visits are then estimated using the following difference-in-differences regression model:

$$\overline{Y}^*_{s,t} = \alpha + \theta_s + \theta_t + \sum_{k=2013}^{2015} \beta_k \cdot \mathbb{1}\{s \in \mathcal{S}_1 \text{ and } t = k\} + \epsilon_{s,t}, \tag{A.12}$$

where θ_s are school fixed effects and θ_t are year fixed effects (using 2012 as the reference year); $\mathbb{1}\{s \in \mathcal{S}_1 \text{ and } t = k\}$ is a dummy variable that take the value one if the observation corresponds to a participating school observed in year k; and $\epsilon_{s,t}$ is the error term. Under the common trend assumption, the coefficient $\hat{\beta}_{2015}$ identifies the average spillover effects among the non-treated students in participating schools. The coefficients $\hat{\beta}_{2013}$ and $\hat{\beta}_{2014}$ provide an indirect test of this assumption: if it holds, the evolution of mean outcomes between 2012 and 2014 (pre-intervention period) should be parallel between participating and non-participating schools, and the coefficients on the pre-interventions "placebos" should not be jointly significant.[A.6]

Selection of non-participating schools. To ensure that non-participating schools are as similar as possible to the participating schools, we use a nearest neighbor matching procedure (with replacement) on the estimated propensity score. We consider all public and private high schools operating in the Paris region that had at least two science track Grade 12 classes in 2015, as this restriction was used in our experimental design to select participating schools (see Section 2 in the main text). We then estimate the probability that the school participated in the experiment in 2015 given a vector of exogenous school characteristics $\mathbf{X}_{st}$ (measured every year between 2012 and 2015) and a vector of the pre-intervention outcomes $\mathbf{Y}_{st}$ (measured in 2012 and 2013) for which spillover effects are measured.[A.7] We then match each participating school with the non-participating school having the closest propensity score among the schools with the same status (public or private) and located in the same education district (Paris, Créteil or Versailles) as that of the participating school.

[A.6] Strictly speaking, the parallel trend assumption only requires the coefficient β_{2013} to be non-statistically significant since, as explained above, the comparison between participating and non-participating schools in 2014 could be contaminated by the classroom interventions that were carried on a small scale that year. As shown below, the results show that the parallel trend assumption also holds between 2013 and 2014, suggesting that the contamination effects of these prior interventions are negligible, if any.

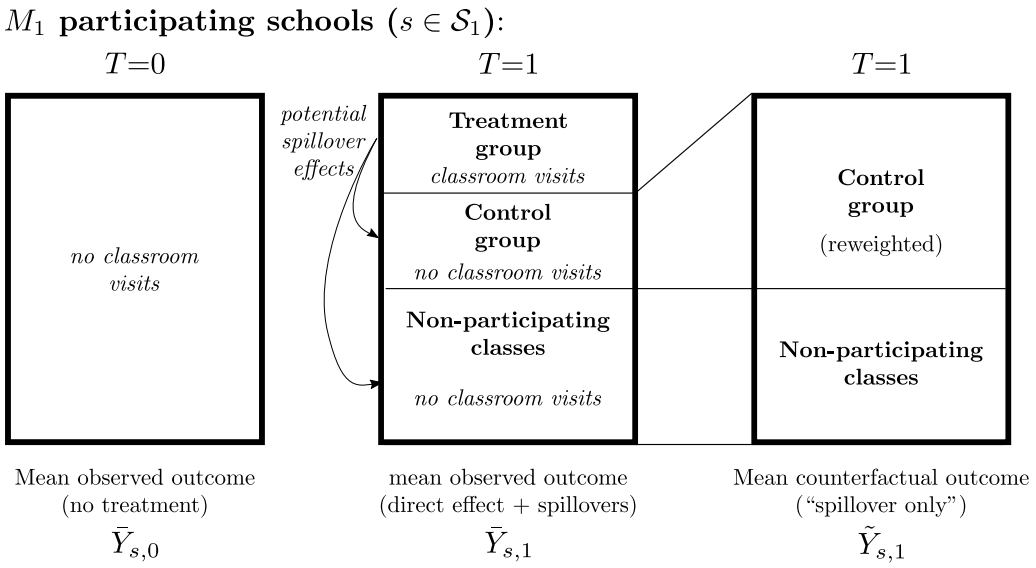
[A.7] The vector of exogenous school characteristics $\mathbf{X}_{st}$ includes the school's education district (Paris, Créteil or Versailles), whether it is public or private, and the following time-varying characteristics every year between 2012 and 2015: the number of students in Grade 12 (science track), the fraction of female students, and the fraction of high-SES students. The vector of pre-intervention outcomes $\mathbf{Y}_{st}$ in 2012 and 2013 includes the fraction of science track Grade 12 students who enroll in a STEM program after graduating from high school, the fraction who enroll in a selective STEM program, and the fraction who enroll in a male-dominated STEM program (computed separately by year and gender). We do not control for pre-intervention outcomes in 2014 to avoid any contamination by classroom interventions that could have been carried out that year.

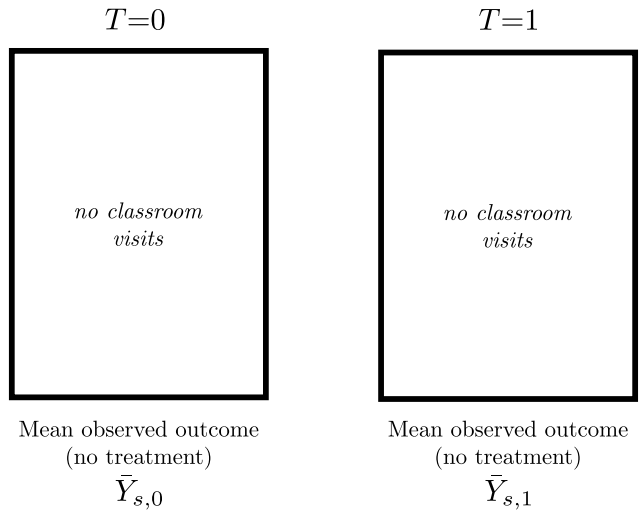
Results. We use Equation (A.12) to estimate the spillover effects of classroom visits on the college enrollment outcomes of Grade 12 students in non-treated classes. The model is estimated separately by gender and we consider the three main outcomes for which we document significant direct effects of the interventions: enrollment in a STEM undergraduate program, enrollment in a selective STEM program, and enrollment in a male-dominated STEM program. The observations are school-by-year averages weighted by school size. Standard errors are clustered at the school level to account for serial correlation across years.

The results are reported in Table K24. Panel A shows that the non-participating schools selected by the nearest-neighbor matching procedure are reasonably similar to the participating schools in terms of the average college enrollment outcomes of female and male students in the pre-intervention period 2012-2013.

The estimates from the DiD regression are reported in Panel B. In all specifications, the coefficients on (participating school $\times$ t=2013) and on (participating school $\times$ t=2014) are close to zero and are neither individually nor jointly significant, which lends support to the assumption of common trends between participating and non-participating schools. Overall, the results provide no evidence of significant spillover effects from the classroom visits in participating schools: for all considered outcomes, the coefficient $\hat{\beta}_{2015}$ on (participating school $\times$ $t = 2015$) is close to zero and not statistically significant for both female and male students.

It should, however, be noted that although our estimates are relatively precise, we cannot rule out small to moderate spillover effects. In the presence of positive spillovers, the treatment effects reported in the main text would under-estimate the true direct effect of classroom visits, since the "contamination" of the control group would push the difference between treatment and control classes towards zero. Denoting by Φ the average direct effect of the classroom interventions and by Δ (> 0) their average indirect effect (through spillovers), the treatment-control difference in mean outcomes, denoted by $\hat{\beta}$, estimates $\Phi - \Delta$ instead of Φ. If we estimate the spillover effects to be at most $\hat{\Delta}^{UB}$, this implies that the size of spillover effects is at most $\hat{\Delta}^{UB}/(\hat{\beta} + \hat{\Delta}^{UB})$ of the size of the direct effect. When we consider the effects on the probability that female students enroll in a selective STEM program, the comparison of treatment and control classes yields an estimated direct effect of $\hat{\beta} = 0.035$ (see Table 6, Panel B, column 2). Based on the results in column 2 of Table K24, the upper bound of the 95 percent confidence interval for the spillover effects is estimated to be $\hat{\Delta}^{UB} = 0.017$. Hence, in the case of selective STEM enrollment, we cannot reject spillover effects that would be at most 33 percent of the size of the "true" direct effect $\hat{\beta} + \hat{\Delta}^{UB}$, which in this case would be of 5.2 percentage points. A similar calculation for the spillover effects on male-dominated STEM enrollment yields an upper bound of $\hat{\Delta}^{UB} = 0.025$. Since the estimated direct effect is $\hat{\beta} = 0.038$, we cannot reject spillover effects of at most 40 percent of the size of the "true'' direct effect $\hat{\beta} + \hat{\Delta}^{UB}$, which in that case would be of 6.3 percentage points.





Difference-in-differences estimator of spillover effects:

$$\hat{\Delta} = \frac{1}{M_1} \sum_{s \in \mathcal{S}_1} (\tilde{Y}_{s,1} - \bar{Y}_{s,0}) - \frac{1}{M_0} \sum_{s \in \mathcal{S}_0} (\bar{Y}_{s,1} - \bar{Y}_{s,0})$$

Figure K7 – Spillover Effects of Role Model Interventions: Empirical Strategy

Notes: This figure illustrates the difference-in-differences strategy we implement to estimate the spillover effects of role model interventions for students who were enrolled in participating schools but whose classes were not assigned to the treatment group. These students are either in the classes that were not selected by school principals to participate in the program evaluation or in the participating classes that were randomly assigned to the control group. Our approach consists in comparing the evolution of mean student outcomes (at the school level) in participating ($s \in \mathcal{S}_1$) and non-participating schools ($s \in \mathcal{S}_0$), between the year before the intervention ($T = 0$) and the year of the intervention ($T = 1$). For $T = 1$, we use a weighted average of non-treated classes in each participating school to estimate the counterfactual "spillover-only" outcome that would have been observed if all the students from that school had only been exposed to the spillover effects of classroom interventions, without being directly exposed to a female role model. Average spillover effects are then estimated by comparing this "spillover-only" counterfactual to a "no-treatment" counterfactual. Under the assumption that absent treatment, mean outcomes in participating school would have followed the same evolution as in non-participating schools, the average spillover effects can be estimated by comparing the evolution between $T = 0$ and $T = 1$ of the mean outcome of students in participating schools (using the "spillover-only" counterfactual for period 1) with the corresponding evolution in non-participating schools.

Table K23 – Scope for Spillover Effects: Summary Statistics from the Student Survey

				Within class	
	All	Boys	Girls	Difference (3)−(2)	p-value of diff.
	(1)	(2)	(3)	(4)	(5)
Panel A. Grade 10					
Treatment Group					
Discussed the classroom visit?					
with classmates	0.580	0.498	0.656	0.145	0.000
with other students from the school	0.240	0.200	0.277	0.072	0.000
with other students outside the school	0.203	0.155	0.247	0.098	0.000
Exposed to other science outreach program?					
this school year	0.128	0.138	0.120	−0.011	0.297
in the past	0.182	0.218	0.149	−0.059	0.000
N	6,245	2,989	3,256		
Control Group					
Heard of classroom visits in other classes?					
Yes, definitely	0.018	0.017	0.020	0.001	0.862
Yes, vaguely	0.122	0.117	0.127	0.009	0.244
No	0.859	0.866	0.853	−0.010	0.271
Exposed to programs about science or jobs in science?					
this school year	0.146	0.144	0.148	0.011	0.283
before the end of this school year	0.052	0.059	0.047	−0.014	0.019
in the past	0.322	0.309	0.333	0.025	0.066
N	5,981	2,762	3,219		
Panel B. Grade 12 (science track)					
Treatment Group					
Discussed the classroom visit?					
with classmates	0.629	0.556	0.705	0.131	0.000
with other students from the school	0.269	0.206	0.334	0.114	0.000
with other students outside the school	0.202	0.133	0.275	0.136	0.000
Exposed to other science outreach programs?					
this school year	0.202	0.200	0.204	0.005	0.797
in the past	0.324	0.349	0.299	−0.053	0.025
N	2,642	1,350	1,292		
Control Group					
Heard of classroom visit in other classes?					
Yes, definitely	0.047	0.049	0.045	−0.004	0.645
Yes, vaguely	0.292	0.275	0.308	0.037	0.048
No	0.661	0.676	0.646	−0.033	0.085
Exposed to programs about science or jobs in science?					
this school year	0.287	0.291	0.284	0.011	0.515
before the end of this school year	0.096	0.104	0.088	−0.009	0.403
in the past	0.488	0.461	0.514	0.054	0.028
N	2,594	1,286	1,308		

Notes: The summary statistics in this table are computed from the post-treatment student survey administered in all treated and control classes between one and six months after the role model interventions. Columns 1, 2, and 3 report average values for all respondents and for boys and girls, respectively, separately by grade level and treatment assignment. The within-class difference in the responses of girls and boys, reported in column 4, is obtained from a regression of the variable of interest on a female dummy, controlling for class fixed effects and clustering standard errors at the school level.

Table K24 – Difference-in-Differences Estimates of the Spillover Effects of Role Model Interventions on College Enrollment Outcomes, Grade 12 Students, Years 2012–2015

	Grade 12 (science track) students					
	Girls			Boys		
	Underg. STEM (1)	Selective STEM (2)	Male-dom. STEM (3)	Underg. STEM (4)	Selective STEM (5)	Male-dom. STEM (6)
Panel A. Baseline means (2012–2013)						
Participating schools						
Mean	0.274	0.145	0.163	0.489	0.265	0.409
Number of schools	88	88	88	87	87	87
Average number of Grade 12 students	107	107	107	108	108	108
Non-participating schools						
Mean	0.265	0.141	0.157	0.473	0.257	0.395
Number of schools	62	62	62	61	61	61
Average number of Grade 12 students	99	99	99	99	99	99
Panel B. Regression estimates						
Pre-trends: participating vs. non-particip. schools (relative to 2012)						
$\hat{\beta}_{2013}$: Particip. school × (t=2013)	0.006	−0.001	0.013	0.003	−0.023	−0.015
	(0.017)	(0.014)	(0.014)	(0.022)	(0.017)	(0.021)
$\hat{\beta}_{2014}$: Particip. school × (t=2014)	0.015	0.001	0.014	0.002	−0.020	−0.017
	(0.019)	(0.014)	(0.014)	(0.018)	(0.015)	(0.017)
Spillover effects: non-treated students						
$\hat{\beta}_{2015}$: Particip. school × (t=2015)	−0.011	−0.014	−0.009	0.008	−0.011	−0.018
	(0.021)	(0.016)	(0.017)	(0.022)	(0.019)	(0.024)
Year fixed effects (omitted: 2012)	Yes	Yes	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations (school×year)	601	601	601	593	593	593
Test: common trends ($\hat{\beta}_{2013}$=$\hat{\beta}_{2014}$=0)						
F-stat	0.33	0.01	0.67	0.01	1.22	0.51
p-value	0.72	0.99	0.52	0.99	0.30	0.60

Notes: This table reports the estimated spillover effects of the role model interventions for students in the non-treated classes of the schools that participated in the program evaluation in 2015, separately for male and female students in Grade 12 (science track). The outcomes we consider are those for which we document significant direct effects of the interventions, i.e., enrollment in a STEM undergraduate program, enrollment in a selective STEM program, and enrollment in a male-dominated STEM program. The results are based on a difference-in-differences specification that compares the outcomes of students in participating and non-participating schools over the period 2012 to 2015, in which the first three years correspond to the pre-intervention period. Non-participating schools are selected among high schools in the Paris region using a nearest neighbor matching procedure (with replacement) on the estimated propensity score. The baseline mean outcomes in participating and non-participating over the pre-intervention period 2012-2013 are reported in Panel A. The regression estimates are reported in Panel B. In all specifications, the dependent variable is the school-by-year average outcome of non-treated students. For non-participating schools throughout the period and for participating schools in the pre-intervention period, this mean outcome is simply the average outcome of all students enrolled in Grade 12 (science track) in the considered year. For participating schools in 2015 (the year of the intervention), this variable is computed as the weighted average outcome of students in the non-participating classes and in the participating classes that were randomly assigned to the control group, with respective weights equal to the share of participating and of non-participating classes (i.e., treatment and control) in the school. The dependent variable is regressed on school fixed effects, year fixed effects (using 2012 as the reference year) and on three dummy variables that take the value one if the observation corresponds to a participating school observed in 2013, 2014, and 2015, respectively. The coefficients on the first two dummy variables capture the differential pre-trends between participating and non-participating schools whereas the coefficient on the third dummy variable measures the spillover effects of role model interventions. All regressions are weighted by school size. Standard errors (in parentheses) are clustered at the school level. The number of schools being used in the regressions for female and male students differs because one of the participating schools and one of the non-participating schools are female-only. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

L Heterogenous Treatment Effects: Subgroup Analysis

Table L25 – Enrollment Status the Following Year, by Level of Performance in Math

	Girls			Boys		
	Treatment effect (LATE) by level of performance in math			Treatment effect (LATE) by level of performance in math		
	Below median	Above median	p-value of diff. [q-value]	Below median	Above median	p-value of diff. [q-value]
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Grade 10						
Grade 11: Science track	−0.019	0.002	0.375	−0.022	0.004	0.358
	(0.015)	(0.019)		(0.021)	(0.018)	
Grade 11: Science - general track	−0.005	0.000	0.834	−0.009	0.013	0.393
	(0.014)	(0.019)	[0.835]	(0.016)	(0.019)	[0.787]
Grade 11: Science - technological track	−0.014**	0.002	0.068	−0.014	−0.009	0.820
	(0.007)	(0.004)	[0.137]	(0.016)	(0.011)	[0.821]
N	3,584	3,484		3,221	3,075	
Panel B. Grade 12 (science track)						
Major: STEM	0.010	0.031	0.571	−0.041	0.016	0.163
	(0.020)	(0.026)		(0.026)	(0.029)	
Major: selective STEM	0.002	0.067***	0.018	−0.014	0.036	0.156
	(0.013)	(0.022)	[0.037]	(0.018)	(0.027)	[0.313]
Major: male-dominated STEM	0.024	0.046**	0.513	−0.005	0.019	0.541
	(0.018)	(0.023)	[0.514]	(0.025)	(0.028)	[0.541]
N	1,544	1,211		1,497	1,328	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' enrollment outcomes in the academic year following the classroom visits, i.e., 2016/17, separately by grade level, gender, and level of academic performance in math. The enrollment outcomes are measured using student-level administrative data. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Students' performance in math is measured from the grades obtained on the final math exam of the *Diplôme national du Brevet* at the end of middle school (for Grade 10 students) and on the final math exam of the *Baccalauréat* (for science track Grade 12 students). Columns 1 and 2 (for girls) and columns 4 and 5 (for boys) report the local average treatment effect (LATE) estimates for students below and above the median level of performance in math, respectively. They are obtained from a regression of the outcome of interest on the interaction between a classroom visit indicator and indicators for the student being below or above the median level of performance in math, using treatment assignment (interacted with the math performance dummies) as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report both the cluster-robust model-based p-value for the difference between the treatment effects estimates for students above vs. below the median performance in math and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table L26 – Enrollment Status the Following Year, by Role Model Background

	Girls			Boys		
	Treatment effect (LATE) by role model background			Treatment effect (LATE) by role model background		
	Researchers	Professionals	p-value of diff. [q-value]	Researchers	Professionals	p-value of diff. [q-value]
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Grade 10						
Grade 11: Science track	0.008	−0.020	0.295	−0.026	0.009	0.272
	(0.020)	(0.018)		(0.024)	(0.019)	
Grade 11: Science - general track	0.016	−0.016	0.225	−0.001	0.008	0.763
	(0.020)	(0.018)	[0.450]	(0.022)	(0.019)	[0.763]
Grade 11: Science - technological track	−0.008	−0.004	0.569	−0.025**	0.001	0.142
	(0.005)	(0.005)	[0.569]	(0.012)	(0.012)	[0.284]
N	3,180	4,061		2,879	3,580	
Panel B. Grade 12 (science track)						
Major: STEM	0.002	0.039**	0.185	−0.007	0.010	0.663
	(0.022)	(0.017)		(0.032)	(0.024)	
Major: selective STEM	0.008	0.053***	0.046	0.008	0.029	0.503
	(0.018)	(0.014)	[0.093]	(0.025)	(0.019)	[0.504]
Major: male-dominated STEM	0.025	0.046***	0.379	−0.002	0.031	0.397
	(0.019)	(0.015)	[0.379]	(0.030)	(0.025)	[0.504]
N	1,180	1,647		1,312	1,612	

Notes: This table reports estimates of the treatment effects of the role model interventions on students' enrollment outcomes in the academic year following the classroom visits, i.e., 2016/17, separately by grade level, student gender, and by background of the female role model who visited the classroom (professional or researcher). The enrollment outcomes are measured using student-level administrative data. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Columns 1 and 2 (for girls) and columns 4 and 5 (for boys) report the local average treatment effect (LATE) estimates for students whose class was visited by a researcher or a professional, respectively. They are obtained from a regression of the outcome of interest on the interaction between a classroom visit indicator and indicators for the role model being either a researcher or a professional, using treatment assignment (interacted with the role model background indicator) as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomization (class). Columns 3 and 6 report both the cluster-robust model-based p-value for the difference between the treatment effects estimates for students visited by a professional vs. a researcher and, in square brackets, the p-value (q-value) adjusted for multiple hypotheses testing across variables belonging to the same family of outcomes, using the False Discovery Rate (FDR) control method. Specifically, we use the sharpened two-stage q-values introduced in Benjamini et al. (2006) and described in Anderson (2008). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table L27 – Grade 12 Students: Enrollment Status the Following Year, by Quintile of Math Performance and by Role Model Background

	Girls			Boys		
	Control group mean	Treatment effect (LATE)	Standard error	Control group mean	Treatment effect (LATE)	Standard error
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. By quintile of math performance at *Baccalauréat*						
Major: STEM	0.289	0.024	0.014	0.470	0.003	0.020
Quintile 1	0.186	−0.020	0.031	0.217	0.044	0.036
Quintile 2	0.282	0.050	0.044	0.441	−0.049	0.041
Quintile 3	0.285	−0.004	0.037	0.573	−0.110	0.052
Quintile 4	0.356	−0.016	0.038	0.570	0.005	0.043
Quintile 5	0.399	0.151	0.045	0.679	0.036	0.046
Major: selective STEM	0.110	0.035	0.011	0.232	0.020	0.016
Quintile 1	0.033	0.000	0.018	0.051	0.011	0.022
Quintile 2	0.040	0.021	0.021	0.127	−0.010	0.026
Quintile 3	0.088	−0.008	0.021	0.242	−0.011	0.046
Quintile 4	0.168	0.038	0.029	0.342	−0.019	0.041
Quintile 5	0.285	0.163	0.043	0.492	0.096	0.047
Major: male-dominated STEM	0.166	0.038	0.012	0.379	0.017	0.019
Quintile 1	0.086	0.008	0.025	0.147	0.061	0.032
Quintile 2	0.141	0.064	0.036	0.333	−0.015	0.038
Quintile 3	0.148	0.025	0.031	0.449	−0.061	0.049
Quintile 4	0.211	−0.001	0.033	0.467	0.013	0.042
Quintile 5	0.311	0.122	0.046	0.617	0.022	0.048
N		2,827			2,924	
Panel B. By role model background						
Major: STEM	0.289	0.024	0.014	0.470	0.003	0.020
Researcher	0.293	0.002	0.022	0.458	−0.007	0.032
Professional	0.285	0.039	0.017	0.480	0.010	0.024
Major: selective STEM	0.110	0.035	0.011	0.232	0.020	0.016
Researcher	0.100	0.008	0.018	0.227	0.008	0.025
Professional	0.118	0.053	0.014	0.236	0.029	0.019
Major: male-dominated STEM	0.166	0.038	0.012	0.379	0.017	0.019
Researcher	0.168	0.025	0.019	0.360	−0.002	0.030
Professional	0.165	0.046	0.015	0.395	0.031	0.025
N		2,827			2,924	

Notes: This table complements Figures 3 and 4 in the main text. It reports estimates of the treatment effects of the role model interventions on Grade 12 (science track) students' enrollment outcomes in the academic year following the classroom visits, i.e., 2016/17, separately by student gender, quintile of performance in math, and role model background (professional or researcher). The enrollment outcomes are measured using student-level administrative data. Students' performance in math is measured from the grades obtained on the final math exam of the *Baccalauréat*. Columns 1 and 4 report the average value for students in the control group. Columns 2 and 5 report the local average treatment effect (LATE) estimates. The estimates shown in Panel A are obtained from a regression of the outcome of interest on interactions between a classroom visit indicator and the quintiles of math performance, using treatment assignment (interacted with the quintiles of math performance) as an instrument for treatment receipt. The estimates shown in Panel B are obtained from a regression of the outcome of interest on interactions between a classroom visit indicator and two indicators for role model type, using treatment assignment (interacted with role model type) as an instrument for treatment receipt. The regression controls for school fixed effects to account for the fact that randomization was stratified by school. Standard errors (reported in columns 3 and 6) are adjusted for clustering at the unit of randomization (classs).

Table L28 – Treatment Effects (ITT) on Enrollment in Selective and Male-dominated STEM among Grade 12 Students: Heterogeneity by Student and Role Model Characteristics

	Dependent variable: undergraduate program is							
	Selective STEM				Male-dominated STEM			
	Girls		Boys		Girls		Boys	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment group indicator (T)	0.001	−0.053*	−0.017	−0.072	0.013	0.020	−0.023	−0.096
	(0.016)	(0.032)	(0.022)	(0.046)	(0.018)	(0.040)	(0.025)	(0.063)
Interactions with student characteristics								
T*Bac rank in math (/100, demeaned)	0.148***	0.162***	0.034	−0.003	0.063	0.070	−0.045	−0.069
	(0.050)	(0.057)	(0.060)	(0.064)	(0.056)	(0.066)	(0.063)	(0.073)
T*Bac rank in French (/100, demeaned)		−0.039		0.091		−0.061		0.012
		(0.045)		(0.060)		(0.060)		(0.068)
T*High SES (demeaned)		0.039		−0.017		0.039		0.034
		(0.028)		(0.032)		(0.031)		(0.039)
Interactions with role model characteristics								
T*Professional	0.060***	0.096***	0.052*	0.076**	0.036	0.086***	0.078**	0.137***
	(0.021)	(0.024)	(0.030)	(0.035)	(0.023)	(0.031)	(0.035)	(0.048)
T*Participated in the program the year before		−0.041		0.026		−0.031		0.012
		(0.025)		(0.040)		(0.023)		(0.044)
T*Age (demeaned)		0.001		0.000		0.001		0.000
		(0.003)		(0.003)		(0.003)		(0.003)
T*Non-French		−0.004		0.003		−0.052***		−0.021
		(0.023)		(0.042)		(0.018)		(0.045)
T*Has children		0.005		0.014		−0.074**		−0.023
		(0.026)		(0.034)		(0.029)		(0.040)
T*Has a Ph.D. degree		0.080***		0.062		0.058*		0.059
		(0.027)		(0.042)		(0.035)		(0.057)
T*STEM Field		−0.035		−0.019		−0.072***		0.051
		(0.025)		(0.033)		(0.025)		(0.040)
Student characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,532	2,532	2,576	2,576	2,532	2,532	2,576	2,576
Adjusted R-squared	0.123	0.123	0.183	0.182	0.062	0.064	0.126	0.125

Notes: Each column corresponds to a separate regression. The sample is restricted to students in Grade 12 (science track). The enrollment outcomes are measured using student-level administrative data. The outcome variable in columns 1–4 (respectively columns 5–9) is an indicator for being enrolled in a selective (respectively male-dominated) STEM undergraduate program in the year following high school graduation, i.e., 2016/17. The coefficients are from a regression of the outcome variable on a treatment group indicator, student characteristics, school fixed effects, and interactions between the treatment group indicator and different subsets of student and role model characteristics. The student characteristics consist of an indicator for high-SES background and percentile ranks on the *Baccalauréat* final exams in math and French. The role model characteristics consist of age and a set of indicators for being a professional, having participated in the program the year before, being non-French, having children, holding a Ph.D. degree, and having graduated from a STEM field. Since each high school was visited by at most one role model, role model fixed effects are absorbed by the school fixed effects. Standard errors (in parentheses) are adjusted for clustering at the class level. The models are estimated separately for girls (odd-numbered columns) and for boys (even-numbered columns). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table L29 – Heterogeneous Treatment Effects on Student Outcomes by Level of Performance in Math and Role Model Background: Grade 10 Students

Panel A. Treatment effects (LATE) by level of performance in math

	Girls			Boys		
	Level of performance in math			Level of performance in math		
	Below median	Above median	p-value of diff. [q-value]	Below median	Above median	p-value of diff. [q-value]
	(1)	(2)	(3)	(4)	(5)	(6)
Grade 11: Science track	−0.019	0.002	0.375	−0.022	0.004	0.358
	(0.015)	(0.019)		(0.021)	(0.018)	
Positive perceptions of science-related careers (index)	0.210***	0.273***	0.281	0.173***	0.154***	0.750
	(0.043)	(0.038)		(0.042)	(0.040)	
More men in science-related jobs	0.169***	0.144***	0.336	0.188***	0.153***	0.148
	(0.019)	(0.017)	[0.448]	(0.020)	(0.017)	[0.224]
Equal gender aptitude for math (index)	0.048	0.168***	0.017	0.098**	0.185***	0.168
	(0.037)	(0.033)	[0.035]	(0.045)	(0.042)	[0.224]
Women like science less than men	0.062***	0.053***	0.688	0.108***	0.096***	0.645
	(0.016)	(0.014)	[0.689]	(0.019)	(0.017)	[0.645]
W face discrimination in science-related jobs	0.171***	0.085***	0.001	0.177***	0.133***	0.111
	(0.019)	(0.017)	[0.004]	(0.020)	(0.019)	[0.224]
Taste for science subjects (index)	−0.072	−0.010	0.274	−0.081**	0.041	0.016
	(0.046)	(0.041)		(0.041)	(0.035)	
Math self-concept (index)	−0.042	0.021	0.232	−0.016	0.080**	0.058
	(0.038)	(0.039)		(0.038)	(0.038)	
Science-related career aspirations (index)	−0.045	0.053	0.079	−0.035	0.044	0.162
	(0.041)	(0.040)		(0.042)	(0.037)	
N	3,584	3,484		3,221	3,075	

Panel B. Treatment effects (LATE) by role model background

	Girls			Boys		
	Role model background			Role model background		
	Researchers	Professionals	p-value of diff. [q-value]	Researchers	Professionals	p-value of diff. [q-value]
	(1)	(2)	(3)	(4)	(5)	(6)
Grade 11: Science track	0.008	−0.020	0.295	−0.026	0.009	0.272
	(0.020)	(0.018)		(0.024)	(0.019)	
Positive perceptions of science-related careers (index)	0.227***	0.258***	0.570	0.136***	0.192***	0.342
	(0.039)	(0.040)		(0.047)	(0.036)	
More men in science-related jobs	0.147***	0.164***	0.512	0.163***	0.173***	0.728
	(0.019)	(0.017)	[0.990]	(0.020)	(0.019)	[0.729]
Equal gender aptitude for math (index)	0.051	0.155***	0.034	0.071	0.209***	0.025
	(0.035)	(0.035)	[0.135]	(0.048)	(0.038)	[0.099]
Women like science less than men	0.055***	0.062***	0.749	0.091***	0.112***	0.399
	(0.017)	(0.014)	[0.990]	(0.018)	(0.018)	[0.532]
W face discrimination in science-related jobs	0.127***	0.127***	0.990	0.135***	0.168***	0.218
	(0.020)	(0.016)	[0.990]	(0.021)	(0.017)	[0.437]
Taste for science subjects (index)	0.017	−0.081*	0.174	−0.093*	0.040	0.043
	(0.054)	(0.048)		(0.048)	(0.042)	
Math self-concept (index)	0.008	−0.020	0.668	0.029	0.048	0.760
	(0.046)	(0.043)		(0.047)	(0.043)	
Science-related career aspirations (index)	0.017	0.007	0.858	−0.028	0.035	0.276
	(0.045)	(0.038)		(0.043)	(0.039)	
N	3,180	4,061		2,879	3,580	

Notes: See notes of Appendix Tables L25 and L26. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table L30 – Heterogeneous Treatment Effects on Student Outcomes by Level of Performance in Math and Role Model Background: Grade 12 Students

Panel A. Treatment effects (LATE) by level of performance in math

	Girls			Boys		
	Level of performance in math			Level of performance in math		
	Below median	Above median	p-value of diff. [q-value]	Below median	Above median	p-value of diff. [q-value]
	(1)	(2)	(3)	(4)	(5)	(6)
Undergraduate major: selective STEM	0.002	0.067***	0.018	−0.014	0.036	0.156
	(0.013)	(0.022)	[0.037]	(0.018)	(0.027)	[0.313]
Undergraduate major: male-dominated STEM	0.024	0.046**	0.513	−0.005	0.019	0.541
	(0.018)	(0.023)	[0.514]	(0.025)	(0.028)	[0.541]
Positive perceptions of science-related careers (index)	0.257***	0.355***	0.277	0.042	0.257***	0.008
	(0.054)	(0.059)		(0.054)	(0.051)	
More men in science-related jobs	0.153***	0.079***	0.050	0.155***	0.144***	0.722
	(0.025)	(0.024)	[0.200]	(0.024)	(0.019)	[0.955]
Equal gender aptitude for math (index)	0.061	0.135***	0.274	0.063	0.211***	0.091
	(0.043)	(0.046)	[0.366]	(0.060)	(0.060)	[0.366]
Women like science less than men	0.028*	0.062***	0.172	0.073***	0.075***	0.954
	(0.015)	(0.016)	[0.345]	(0.023)	(0.021)	[0.955]
W face discrimination in science-related jobs	0.116***	0.088***	0.489	0.090***	0.050*	0.368
	(0.027)	(0.030)	[0.490]	(0.030)	(0.028)	[0.736]
Taste for science subjects (index)	−0.054	0.025	0.342	−0.034	0.016	0.553
	(0.051)	(0.056)		(0.058)	(0.052)	
Math self-concept (index)	0.061	−0.070	0.084	0.078*	0.032	0.488
	(0.051)	(0.053)		(0.046)	(0.045)	
Science-related career aspirations (index)	0.061	0.137**	0.353	0.008	0.060	0.514
	(0.049)	(0.060)		(0.054)	(0.050)	
N	1,544	1,211		1,497	1,328	

Panel B. Treatment effects (LATE) by role model background

	Girls			Boys		
	Role model background			Role model background		
	Researchers	Professionals	p-value of diff. [q-value]	Researchers	Professionals	p-value of diff. [q-value]
	(1)	(2)	(3)	(4)	(5)	(6)
Undergraduate major: selective STEM	0.008	0.053***	0.046	0.008	0.029	0.503
	(0.018)	(0.014)	[0.093]	(0.025)	(0.019)	[0.504]
Undergraduate major: male-dominated STEM	0.025	0.046***	0.379	−0.002	0.031	0.397
	(0.019)	(0.015)	[0.379]	(0.030)	(0.025)	[0.504]
Positive perceptions of science-related careers (index)	0.197***	0.386***	0.005	0.151***	0.158***	0.912
	(0.055)	(0.041)		(0.045)	(0.047)	
More men in science-related jobs	0.150***	0.110***	0.213	0.158***	0.142***	0.608
	(0.026)	(0.021)	[0.445]	(0.023)	(0.020)	[0.608]
Equal gender aptitude for math (index)	0.124***	0.077**	0.422	0.201***	0.078	0.128
	(0.047)	(0.035)	[0.563]	(0.063)	(0.051)	[0.513]
Women like science less than men	0.045***	0.044***	0.931	0.088***	0.062***	0.357
	(0.014)	(0.012)	[0.931]	(0.024)	(0.017)	[0.608]
W face discrimination in science-related jobs	0.126***	0.076***	0.222	0.083***	0.064***	0.581
	(0.034)	(0.024)	[0.445]	(0.028)	(0.022)	[0.608]
Taste for science subjects (index)	−0.044	0.055	0.152	−0.014	0.010	0.750
	(0.054)	(0.044)		(0.056)	(0.052)	
Math self-concept (index)	0.108*	0.013	0.231	0.173***	−0.006	0.010
	(0.060)	(0.051)		(0.055)	(0.044)	
Science-related career aspirations (index)	−0.093	0.246***	0.000	0.028	0.068	0.546
	(0.057)	(0.044)		(0.052)	(0.042)	
N	1,180	1,647		1,312	1,612	

Notes: See notes of Appendix Tables L25 and L26. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

M Heterogeneous Treatment Effects: Machine Learning Methods

This appendix provides additional information on the machine learning methods we use to (i) describe the heterogeneity in treatment effects and (ii) estimate the correlation between treatment effects on different outcomes. Section M.1 gives an overview of the generic approach recently developed by Chernozhukov et al. (2018) to estimate, and make inference about, key features of heterogeneous effects in randomized experiments. Section M.2 provides further details on how we implement this method in the context of our study. Section M.3 explains how we extend this method to estimate the correlation between treatment effects, which we view as a promising alternative to causal mediation analysis when exploring channels of influence in RCTs. Finally, Section M.4 draws attention to the fact that the adjusted p-values proposed by Chernozhukov et al. (2018) lead to conservative inference.

M.1 Description of the Method of Chernozhukov et al. (2018)

Motivation. Reporting treatment effects for various subgroups of participants opens the possibility of overfitting due to the large number of potential sample splits. To address this issue, one option is to specify a certain number of groups *ex ante* in a pre-analysis plan and to tie one's hands to analyze treatment effect heterogeneity only across these groups, while correcting standard errors for multiple testing. This approach, however, has the drawback of restricting the analysis to a small number of groups and bears the risk of missing important sources of heterogeneity.

Machine Learning (ML) methods provide an attractive alternative to explore treatment effect heterogeneity in a more comprehensive manner (see Athey and Imbens, 2017, for a review). We adopt the approach developed by Chernozhukov et al. (2018) as it appears well-suited for our objective. First, this approach makes it possible to conduct valid statistical inference on several objects of interest, such as average treatment effects by heterogeneity groups or the characteristics of individuals with large and small predicted treatment effects. Second, this method can be implemented using any ML method, allowing us to test algorithms of different degrees of sophistication, ranging from simple linear models to neural networks. Third, as described in Section M.3, this approach can be extended to estimate the correlation between treatment effects on different outcomes.

Concepts and estimation procedure. Consider an outcome variable denoted by Y. Let $Y(1)$ and $Y(0)$ denote the potential outcomes of a student when her class is and is not visited by a role model, respectively. Let Z be a vector of covariates that characterize the student and the role model who visited the class. The conditional average treatment effect (CATE), denoted by $s_0(Z)$, is defined as follows:

$$s_0(Z) \equiv \mathbb{E}[Y(1) - Y(0)|Z].$$

The approach developed by Chernozhukov et al. (2018) uses the following procedure:

1. Randomly split the data into a *training sample* and an *estimation sample* of equal size (using stratified splitting to balance the proportions of treated and control units in each subsample).
2. Use the training sample to predict the CATE using various ML algorithms. Obtain a ML predictor proxy predictor $S(Z)$.
3. Estimate and perform inference on *features* of the CATE on the estimation sample (see the definition of the features below).

4. Repeat steps 1 to 3 n times and keep track of the estimates obtained for each feature as well as their associated p-values and 95 percent confidence intervals.
5. For each feature, compute the final estimate as the median of the n available estimates. Compute the p-value for this final estimate as the median of the n available p-values multiplied by two. Compute a 90 percent confidence interval for the final estimate as the median of the n 95 percent confidence intervals.

Three features of the CATE. The CATE $s_0(Z)$ is a function for which it is difficult to obtain uniformly valid inference without making strong assumptions. It is, however, possible to obtain inference results on specific *features* of the CATE, such as the expectation of $s_0(Z)$ for heterogeneity groups induced by the ML proxy predictor $S(Z)$.

The Best Linear Predictor (BLP). The first feature of the CATE $s_0(Z)$ is its Best Linear Predictor (BLP) based on the ML proxy predictor $S(Z)$. It is formally defined as follows:

$$\text{BLP}[s_0(Z)|S(Z)] \equiv \underset{f(Z)\in \text{Span}(1,S(Z))}{\arg\min} \mathbb{E}[s_0(Z) - f(Z)]^2.$$

Chernozhukov et al. (2018) show that one can identify the BLP of $s_0(Z)$ given $S(Z)$, as well as the projection parameters $\beta_1 = \mathbb{E}[s_0(Z)]$ and $\beta_2 = \text{Cov}(s_0(Z), S(Z))/\text{Var}(S(Z))$, using the following weighted linear projection:

$$Y = \alpha_0 + \alpha B(Z) + \beta_1(T - p(Z)) + \beta_2(T - p(Z))(S(Z) - \mathbb{E}[S(Z)]) + \epsilon, \;\; \mathbb{E}[w(Z)\epsilon X] = 0, \quad \text{(A.13)}$$

where T is the treatment group indicator; $B(Z)$ is a ML predictor of $Y(0)$ obtained from the training sample; $p(Z)$ is the propensity score (i.e., the conditional probability of being assigned to the treatment group); $w(Z) \equiv \{p(Z)(1 - p(Z))\}^{-1}$ is the weight; and X is the vector of all regressors ($X \equiv [1, B(Z), T - p(Z), (T - p(Z))(S(Z) - \mathbb{E}[S(Z)])]$).

Equation (A.13) can be estimated using weighted least squares, after replacing $\mathbb{E}[S(Z)]$ by its empirical expectation with respect to the estimation sample.

The coefficient β_2 is informative about the correlation between the true CATE, $s_0(Z)$ and the predicted CATE, $S(Z)$. It is equal to one if the prediction is perfect and to zero if $S(Z)$ has no predictive power or if there is no treatment effect heterogeneity, that is $s_0(Z) = s$. The main purpose of estimating β_2 is to check if the trained ML algorithms are able to detect heterogeneity.[A.8]

Sorted Group Average Treatment Effects (GATEs). The ML predictor of the CATE, $S(Z)$, can be used to identify groups of individuals with small and large predicted treatment effects. In our setting, this is achieved by sorting students in the estimation sample (indexed by i) according to $S(Z_i)$, the predicted value of their treatment effect given their observable characteristics. We consider the bottom and top quintiles of $S(Z_i)$ and provide ITT estimates for both groups of students.

Classification Analysis (CLAN). The third feature consists in comparing the distribution of observable characteristics of students with the smallest and largest predicted treatment effects.

The three above features—the BLP, the GATEs, and the CLAN—all rely on the existence of a ML predictor $S(Z)$. The BLP provides a means to check if $S(Z)$ detects significant heterogeneity in treatment effects. If it fails to do so, the GATEs and CLAN are not particularly

[A.8] The intuition behind the formula for β_2 can be grasped by noting that Equation (A.13) is a variant of the simpler equation $Y = \alpha_0 + \alpha B(Z) + \beta_2' T \cdot S(Z) + \epsilon$. This simpler model implies that $s_0(Z) = \beta_2' S(Z)$, suggesting that β_2' provides an estimate for how close the machine learning predictor $S(Z)$ is to the CATE $s_0(Z)$.

relevant for the analysis, as these features would provide a description of students for whom the predicted treatment effect only differs from the unobserved CATE because of a poor-quality prediction.

M.2 Implementation of the Method

This section provides details on the implementation of the method of Chernozhukov et al. (2018) in our empirical setting.

Populations of interest. In the main text, we focus on the sample of girls in Grade 12 (science track), since this group of students is the only one for which we find significant treatment effects on enrollment outcomes. We identify which of these female students were most affected by the program and investigate the messages to which they were particularly responsive. Results for boys in Grade 12 can be found in Appendix Table M33.

Sample splits and iterations. We perform $n = 100$ iterations of the procedure described in the previous section, which consists in (i) splitting the sample into a training and an estimation subsample of equal size; (ii) predicting the CATE on the training sample using ML methods; and (iii) estimating the three features of the CATE (BLP, GATEs, and CLAN) in the estimation sample.[A.9] The sample splits are stratified by class, which is the randomization unit in our experimental setting: half of the girls in each Grade 12 class are randomly assigned to the training sample, while the other half are assigned to the estimation sample.

Propensity score. For each student, we estimate the probability that his or her class was randomly assigned to the treatment group. This propensity score $p(Z)$ is equal to one half in most cases as the treatment was generally assigned to two Grade 10 classes out of four and to one Grade 12 class out of two among the classes that were selected by the school principals. In other cases, the propensity score is not exactly one half.

Machine learning methods. We consider five alternative machine learning methods to estimate the proxy predictor $S(Z)$: Elastic Net, Random Forest, Boosted Trees, Neural Network with feature extraction, and a simple linear model estimated via OLS. These methods are implemented in R using the `caret` package written by Kuhn (2008), while the general approach of Chernozhukov et al. (2018) is implemented by adapting the codes made available online by the authors.[A.10]

For each machine learning method, the predictor $S(Z)$ is constructed in several steps. First, the model is fitted separately on the treatment and control group students in the training sample. The two fitted models are then applied to the estimation sample to obtain the predicted outcomes $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ for each individual. Finally, $S(Z)$ is obtained by taking the difference between the two predictions.[A.11]

For each outcome, we estimate the BLP of the CATE based on the ML method whose associated predictor $S(Z)$ has the highest correlation with the CATE $s_0(Z)$ in the estimation

[A.9] The medians of the estimated features of the CATE change little when we repeat the entire procedure using a different seed number to randomly split the data into the training and estimation samples, suggesting that 100 iterations are sufficient for the purpose of empirical convergence.

[A.10] `https://github.com/demirermert/MLInference` (accessed on May 4, 2018).

[A.11] Predicting outcomes for treatment and control individuals separately, before taking the difference as we do here may not be the most efficient approach to predict the CATE at finite distance. In our setting, however, alternative ML methods directly designed to detect heterogeneity in treatment effects, such as the causal forests proposed by Wager and Athey (2018), did not improve performance. We therefore decided not to rely on these ML methods for the main analysis.

sample. In practice, the best ML method for the BLP targeting of the CATE is chosen in the estimation sample by maximizing the following performance measure:

$$\Lambda \equiv |\beta_2|^2 \text{Var}(S(Z)) = \text{Corr}^2(s_0(Z), S(Z))\text{Var}(s_0(Z)).$$

The above equation shows that maximizing Λ is equivalent to maximizing the correlation between the ML predictor $S(Z)$ and the CATE $s_0(Z)$.

The best method for the GATEs targeting of the CATE, and hence also for the CLAN, is selected based on the following performance measure:

$$\overline{\Lambda} \equiv \mathbb{E}\left(\sum_{k=1}^{K} \gamma_k \mathbb{1}(S \in I_k)\right)^2,$$

where K is the number of (equal-sized) heterogeneity groups, $I_k = [l_{k-1}, l_k)$ are non-overlapping intervals that divide the support of S into regions $[l_{k-1}, l_k)$ with equal or unequal masses, and γ_k is the GATE for heterogeneity group k. In practice, both performance measures lead to a similar ranking of ML methods and the methods eventually selected to produce the BLP, the GATEs/CLAN are almost always the same.

Predictors. The covariates we use to train the ML methods for the samples of boys and girls in Grade 12 are three indicators for the education districts of Paris, Créteil, and Versailles, four indicators for students' socio-economic background (high SES, medium-high SES, medium-low SES, and low SES), their age, their overall percentile rank in the *Baccalauréat* exam, their percentile ranks in the French and math tests of the exam, and a vector of 56 role model fixed effects.[A.12] Our motivation for including only a few pre-determined covariates in addition to the role model indicators is that we are mostly interested in the treatment effect heterogeneity that arises from the 56 role models (which can be seen as 56 treatment arms).

M.3 Correlation Between Treatment Effects on Different Outcomes

In this section, we explain how the method of Chernozhukov et al. (2018) can be extended to estimate the correlation between the treatment effects on different outcomes. We show that a set of four linear projections of the CATEs for two outcomes Y^A and Y^B on the ML predictors of the CATEs for these outcomes can be combined to estimate the correlation between the two CATEs under a natural assumption about prediction errors. We argue that this approach offers a promising alternative to other methods, such as causal mediation analysis, that are commonly used in the medical and social sciences literature to identify what factors may be part of the causal pathway between an intervention and an outcome. Indeed, our proposed method does not rely on strong identifying assumptions and can be used in any experimental setting, as long as there is a sufficiently large number of observed exogenous covariates.

A new feature: projecting a CATE on the predictor of another CATE. Let Y^A and Y^B denote two distinct outcomes and let $s_0^A(Z)$ and $s_0^B(Z)$ denote the true CATEs of a treatment T on these outcomes, given a vector of exogenous covariates Z characterizing the observational units (indexed by i). Let $\rho_{A,B|Z} \equiv \text{Corr}(s_0^A(Z), s_0^B(Z))$ denote the bivariate

[A.12] Each student in the control group is assigned to the role model who visited his or her high school to ensure that the role model indicators are defined for students in both the treatment and control groups. Moreover, to account for the fact that some Grade 12 students have missing *Baccalauréat* grades (less than 2 percent), we include indicators for missing grades as controls.

correlation between the CATEs on Y^A and Y^B and consider the following weighted linear projection:

$$Y^A = \alpha_0 + \alpha B^B(Z) + \beta_1(T - p(Z)) + \beta_2(T - p(Z))(S^B(Z) - \mathbb{E}[S^B(Z)]) + \epsilon, \;\; \mathbb{E}[w(Z)\epsilon X] = 0, \tag{A.14}$$

where $B^B(Z)$ and $S^B(Z)$ are a ML predictor of outcome Y^B for individuals in the control group and a ML predictor of the CATE on Y^B, respectively. Both ML predictors are trained using a separate independent sample and are taken as given functions in Equation (A.14). The functions $p(Z)$ and $w(Z)$ and the vector X have the same meaning as in Equation (A.13). Equation (A.14) is estimated using weighted least squares, after replacing $\mathbb{E}[S^B(Z)]$ by its empirical expectation with respect to the estimation sample.

Adapting the BLP equation of Chernozhukov et al. (2018) (Equation 2.1 p. 8) by replacing the ML predictor of the CATE on outcome Y^A by the ML predictor of the CATE for outcome Y^B, we directly obtain that Equation (A.14) identifies

$$\beta_2^{A|B} = \text{Cov}(s_0^A(Z), S^B(Z))/\text{Var}(S^B(Z)).$$

The sign of $\beta_2^{A|B}$ is informative of the extent to which the CATE on Y^A is positively or negatively correlated with the CATE on Y^B. To show this formally, we denote by η_B the approximation error in $S^B(Z)$ and we write $S^B(Z) = s_0^B(Z) + \eta_B$. Assuming that η_B is independent of $s_0^A(Z)$, we get that $\beta_2^{A|B} = \text{Cov}(s_0^A(Z), s_0^B(Z))/\text{Var}(S^B(Z))$, which implies that $\beta_2^{A|B}$ and $\rho_{A,B|Z}$ have the same sign.

Combining BLPs to recover the correlation between treatment effects. For any pair of indices $(k, l) \in \{(A, A), (B, B), (A, B), (B, A)\}$, we can identify

$$\beta_2^{k|l} = \text{Cov}(s_0^k(Z), S^l(Z))/\text{Var}(S^l(Z))$$

from the BLP of $s_0^k(Z)$ on $S^l(Z)$. Writing $S^A(Z) = s_0^A(Z) + \eta_A$, $S^B(Z) = s_0^B(Z) + \eta_B$, and assuming that the prediction errors η_A and η_B are independent of both the predicted functions $s_0^A(Z)$ and $s_0^B(Z)$ in the estimation sample, we can write

$$\beta_2^{k|l} = \text{Cov}(s_0^k(Z), s_0^l(Z))/(\text{Var}(s_0^l(Z)) + \text{Var}(\eta^l(Z))).$$

Combining the formulas for the four different possible BLPs, we obtain the following expression:

$$\rho^2_{A,B|Z} = \frac{\beta_2^{A|B}\beta_2^{B|A}}{\beta_2^{B|B}\beta_2^{A|A}},$$

which implies that the correlation $\rho_{A,B|Z}$ is identified as

$$\rho_{A,B|Z} = \text{Sign}(\beta_2^{A|B}) \frac{\sqrt{\beta_2^{A|B}\beta_2^{B|A}}}{\sqrt{\beta_2^{B|B}}\sqrt{\beta_2^{A|A}}}. \tag{A.15}$$

Practical implementation. As explained in the main text, we use the method of Chernozhukov et al. (2018) to estimate the four heterogeneity loading parameters $\beta_2^{A|A}$, $\beta_2^{B|B}$, $\beta_2^{A|B}$, and $\beta_2^{B|A}$. At each iteration of the data-splitting process, the bivariate correlation $\rho_{A,B|Z}$ is estimated by plugging the four parameter estimates into Equation (A.15). In theory, $\beta_2^{A|A}$ and $\beta_2^{B|B}$ should both be positive while $\beta_2^{A|B}$ and $\beta_2^{B|A}$ should have the sign of $\rho_{A,B|Z}$ in each iteration of the data-splitting process. However, it can happen that the estimates $\hat{\beta}_2^{A|A}$, $\hat{\beta}_2^{B|B}$, $\hat{\beta}_2^{A|B}$, and

$\hat{\beta}_2^{B|A}$ do not satisfy these conditions due to estimation error, in particular when the predictors $S^A(Z)$ and $S^B(Z)$ are very noisy. In such cases, we do not estimate $\rho_{A,B|Z}$ and discard the corresponding iteration of the data-splitting procedure. We iterate until we reach a number of 100 iterations for which $\hat{\rho}_{A,B|Z}$ can be computed, so that our final estimates are medians computed over an identical number of iterations.[A.13]

The estimates based on Equation (A.15) can become very large (well above one in absolute value) when the estimates of $\hat{\beta}_2^{A|A}$ or $\hat{\beta}_2^{B|B}$ are close to 0, which can occur when either or both of the predictors $S^A(Z)$ and $S^B(Z)$ are noisy. Reassuringly, we show in Appendix Table M35 that the correlation estimates $\hat{\rho}_{A,B|Z}$ are hardly affected when we exclude data splits that yield a poor ML prediction of the CATEs on outcomes Y^A or Y^B, by using only the first 100 iterations of the data-splitting process for which the estimates of the heterogeneity loading parameters $\hat{\beta}_2^{A|A}$ and $\hat{\beta}_2^{B|B}$ are above a minimum threshold t.

In the absence of a closed-form formula for the standard error of $\hat{\rho}_{A,B|Z}$, we estimate its 95 percent confidence interval as follows. At each iteration m of the data-splitting process, we compute $\hat{\rho}_{A,B|Z}^{(m)}$ (indexed by m) in the estimation sample. When $\hat{\rho}_{A,B|Z}^{(m)}$ can be computed, we estimate its 97.5 % confidence interval using a clustered bootstrap procedure, which accounts for clustered nature of the treatment assignment (at the class level). This procedure consists in creating B replications of the estimation sample m by drawing with replacement $N_c^{(m)}$ female students from each Grade 12 class c, where $N_c^{(m)}$ is the number of female students from class c in the estimation sample m, and computing $\rho_{A,B|Z}$ for this bootstrap sample. For each estimation sample m, this operation is repeated 6,000 times to estimate the 97.5 percent confidence interval of $\hat{\rho}_{A,B|Z}^{(m)}$ using the bootstrap percentile confidence interval method (Davison and Hinkley, 1997, chap. 5).[A.14] The 95 percent confidence interval for $\hat{\rho}_{A,B|Z}$ is then computed as the median of the 97.5 percent confidence intervals over the first 100 iterations for which $\hat{\rho}_{A,B|Z}^{(m)}$ could be computed—the price of the splitting uncertainty being reflected in the discounting of the confidence level from $1-\alpha$ to $1-2\alpha$.

Comparison with alternative strategies to correlate treatment effects. A simpler and perhaps more intuitive approach to evaluate if the role models who had the largest effects on enrollment outcomes also had larger effects on other outcomes would be to estimate separate treatment effects for each role model and to compute the correlation between these role-model-specific treatment effects for outcomes Y^A and Y^B. Unfortunately, this approach would likely result in a biased estimate of the true correlation between treatment effects due to sampling error. Indeed, although the random assignment of treatment ensures that the characteristics of treatment and control students are balanced in the overall sample, this will not in general be true in each of the smaller subsamples of students who were visited by a given role model. In other words, these subsamples may not be entirely comparable to their control group counterparts in the absence of the interventions. The role-model-specific treatment effects for outcomes Y^A and Y^B are therefore likely to be contaminated by imbalances between the treatment and control students attached to each role model. The problem stems from the fact that as long as the outcomes Y^A and Y^B are correlated, baseline differences between control and treated students in terms of these outcomes will also be correlated, and so will be the correlation between the estimated fixed effects.[A.15] To address this issue, one would need to use different samples of

[A.13] For each pair of outcomes (Y^A, Y^B), Appendix Table M34 indicates the proportion of random data splits for which the correlation between CATEs could be computed.

[A.14] The 97.5 percent confidence interval of $\hat{\rho}_{A,B|Z}^{(m)}$ is estimated using only the bootstrap samples for which $\hat{\rho}_{A,B|Z}$ can be computed.

[A.15] To give an example, suppose that a given role model visited classes in which students' perceptions of science-related careers were substantially more positive at baseline than in the school's control group classes. In this scenario, students in the visited classes would be expected to have a higher probability of enrolling in

students to estimate the role model fixed effect for outcomes Y^A and Y^B, which is close in spirit to the method we propose.

Although our approach is not subject to the fundamental bias described above, it has the limitation of only exploiting variation in treatment effects arising from variation in observable characteristics Z. If treatment effects vary more with unobserved characteristics than with Z, $\rho_{A,B|Z}$ will miss much of the existing heterogeneity in treatment effects. Formally, this means that $\rho_{A,B|Z}$ is likely to differ substantially from the true correlation between individual-level treatment effects, $\rho_{A,B|i}$ if the vector of covariates Z is too limited to adequately capture the treatment effect heterogeneity. A direct identification of $\rho_{A,B|i}$ would avoid this limitation. Our investigations suggest, however, that identifying this parameter requires strong and potentially unrealistic assumptions.[A.16]

Comparison with mediation analysis. Causal mediation analysis is extensively conducted by applied researchers in psychology, as well as in other fields such as biomedical science, political science, and sociology (see MacKinnon et al., 2007, for a survey). This approach is used to identify the channels through which a treatment affects an outcome, using a variety of methods ranging from linear structural equation models (Baron and Kenny, 1986) to less parametric identification of causal mechanisms (Imai et al., 2010). Many authors, however, have drawn attention to the fact that in the absence of convincing instruments (as in Frölich and Huber, 2017) or of multiple treatment arms (as in Imai et al., 2013), strong identifying assumptions are required. These assumptions are not always explicitly stated by applied researchers (see Heckman et al., 2013 and Keele, 2015, for counter-examples in the field of economics), although they can often be challenged in the particular context under study.

The reason why strong assumptions are required in causal mediation analysis is relatively straightforward. Randomization ensures that the experimenter can examine the effect of a treatment on any variable of interest, including a *final outcome* and possible mediators (*channels of influence*, in our terminology). What is missing is the causal effect of the mediators on the outcome. In the absence of a clear understanding of how a mediating variable affects the outcome, it is difficult to convincingly conclude that the treatment effect on the final outcome can be attributed to its effect on the mediator.

Our approach differs from mediation analysis in that it does not rely on the observed mapping between the mediators and the outcome of interest, and hence does not require the researcher to interpret this mapping as causal. This key distinction between the two approaches can be easily understood in the context of our study. We observe among girls in Grade 12 a positive correlation between the belief that women are underrepresented in science-related careers and their future enrollment in selective STEM undergraduate programs (the correlation is 0.12 in the control group and 0.75 in the treatment group). This positive correlation could be explained by the fact that girls considering STEM studies are better informed about the underrepresentation of women in STEM fields. However, for causal mediation analysis to be feasible, one would need to give a causal interpretation to this correlation and assume that *increasing* girls' awareness

STEM studies than students in the control group classes, independently of the role model's treatment effect. Hence the role model's larger fixed effect on both outcomes would spuriously capture the positive correlation between baseline outcomes, which has nothing to do with the correlation between treatment effects.

[A.16] A first route would be to assume that the treatment is *rank preserving* for both outcomes Y^A and Y^B, meaning that the intervention does not affect the rank of students with respect to the considered outcome variable in the treatment and control groups. Under this assumption, one could retrieve individual-level treatment effects for both Y^A and Y^B, and compute the correlation between them. In our setting, this assumption seems difficult to justify. An alternative route would be to recover $\rho_{A,B|i}$ from the treatment effect on the *product* of outcomes Y^A and Y^B, in the spirit of Dupas et al. (2018). Unfortunately, this strategy can be shown to identify $\rho_{A,B|i}$ only under restrictive conditions (details available upon request). By contrast, our proposed approach has the advantage of not relying on strong identifying assumptions.

of female underrepresentation in STEM steers them towards STEM studies, which would be a highly questionable assumption in our setting. As a matter of fact, it would lead us to conclude that part of the effect of role models on selective STEM enrollment is mediated by the fact that they tend to make female underrepresentation in STEM more salient, which is at odds with what our analysis suggests. Indeed, we find that the girls for whom the program had the largest effects on selective STEM enrollment are, rather, those whose awareness of the underrepresentation of women in STEM increased the least. This example illustrates the difficulties in drawing robust conclusions on causal mechanisms using standard mediation analysis. By contrast, our approach, which limits itself to evaluating which outcomes are jointly affected by the treatment, appears more agnostic and transparent. Although not causal, the correlations that it uncovers can provide useful hints on the channels of influence.

M.4 Interpretation of Inference Results in Chernozhukov et al. (2018)

Using different samples for prediction and estimation is a key component of machine learning methods to avoid over-fitting. It also makes it possible to carry out standard inference. Provided that the sample splitting is random, data analysis on the estimation sample can be performed using standard statistical techniques and leads to valid inference when the ML predictor is trained on a distinct sample. The novelty introduced by Chernozhukov et al. (2018) is to iterate the data-splitting process and to take the medians of estimates, p-values, and confidence intervals over multiple splits. This approach limits the risk of data mining around alternative data splits, and yields estimates that are more representative of the whole sample.[A.17] This improvement, however, comes at a cost, as the medians of the p-values over the data splits have to be multiplied by two for the inference to be valid.

In this section, we emphasize that the p-values (and the related confidence intervals) proposed by Chernozhukov et al. (2018) should be interpreted as upper bounds of the true p-values. We argue that the conditions under which these bounds are reached are unlikely to be met by practitioners using real data. In the absence of alternative approaches to conduct inference, we systematically report in the main text the adjusted p-values advocated by Chernozhukov et al. (2018), as they have well-defined properties. However, based on the arguments below, we consider them as conservative.

Why should adjusted *p*-values be interpreted as upper bounds? The inference approach proposed in Section 4 of Chernozhukov et al. (2018) relies on the mathematical properties of uniformly distributed random variables to quantify the uncertainty coming from both parameter estimation and data splitting. Indeed, under the null hypothesis, the p-value of a test statistic is uniformly distributed over the interval $[0, 1]$.[A.18] This is true in particular for the p-values of the estimates obtained on the estimation sample after each sample split. Hence, to provide valid inference over n sample splits using the median of the p-values obtained for each split, it is necessary to understand how this median is distributed. Chernozhukov et al. (2018) rely on the following property of uniform variables:

Lemma 3.1 (Chernozhukov et al., 2018). *Consider M_n, the (lower) median of a sequence $\{U_j\}_{j=1}^n$ of n uniformly distributed variables, $U_j \sim U(0,1)$ for each j, where the variables are*

[A.17] To put it differently, estimates from a single random split of the data may yield misleading conclusions by chance, which cannot happen when the final estimates are computed as medians of estimates from multiple splits.

[A.18] Under the null hypothesis, the p-value has a probability α of being lower than α, implying that the null hypothesis has a probability α to be falsely rejected at the significance level α.

not necessarily independent. Then, for $\alpha \in [0,1]$,

$$\mathbb{P}(M_n \leq \alpha) \leq 2\alpha. \tag{A.16}$$

Without entering the details of the proof of the inference result, the above lemma helps to understand why the median of the p-values has to be multiplied by two to yield a valid p-value for the median of the estimates. Under the null hypothesis, the median p-value can have a probability as high as 2α to be lower than α, implying that the null hypothesis may be rejected with probability 2α even if the median p-value is α.

Note, however, that the inequality (A.16) only provides an upper bound for the true p-value, which is given by $\mathbb{P}(M_n \leq \alpha)$. If, for a specific sequence of uniformly distributed variables U_j, one can obtain a tighter upper bound α' for $\mathbb{P}(M_n \leq \alpha)$, p-values would only need to be multiplied by α'/α instead of 2.[A.19]

Why are the upper bounds likely to be conservative? To get a sense of how conservative the upper bound for $\mathbb{P}(M \leq \alpha)$ given by the inequality (A.16) might be, we discuss two polar cases in which this bound is not reached, and one example in which it is.

Polar case 1: Independent uniform variables. Assume that the variables U_j are independent from each other. Define $B_j \equiv \mathbb{P}(U_j \leq \alpha)$. Then $\mathbb{P}(M_n \leq \alpha) = \mathbb{P}(\sum_{j=1}^n B_j/n \geq 1/2)$. If the sequence $\{U_j\}_{j=1}^n$ is i.i.d., the sequence $\{B_j\}_{j=1,...,n}$ is i.i.d. as well. Applying the law of large numbers for i.i.d. variables, we get that $\sum_{j=1}^n B_j/n \xrightarrow{n\to\infty} \alpha$, which implies that $\mathbb{P}(M_n \leq \alpha) \xrightarrow{n\to\infty} 0$ for $\alpha < 1/2$.

Using Hoeffding's inequality, we can also show that

$$\mathbb{P}\left(\sum_{j=1}^n B_j/n \geq 1/2\right) \leq \exp(-2n(1/2-\alpha)^2),$$

which implies that $\mathbb{P}(M_n \leq \alpha)$ converges to 0 at an exponential rate.

Polar case 2: Perfectly positively correlated uniform variables. Assume instead that the random variables U_j are perfectly positively correlated so that their realizations are all equal. In that case, $\mathbb{P}(M_n \leq \alpha) = \mathbb{P}(U_1 \leq \alpha) = \alpha$.

These two polar cases might suggest that depending on the magnitude of the positive correlation between the uniform variables, $\mathbb{P}(M_n \leq \alpha)$ lies between 0 and α. This is, however, not true in the general case, as shown by the following counter-example.

An example in which the upper bound is attained. We now provide an example in which the bound defined in the inequality (A.16) is attained when n is an even number.

Consider a given $\alpha < 1/2$ and denote $k = n/2$. Let V and V' be two independent uniform variables in $[0,1]$, and $A \subset \{1,2,...,n\}$ a random subset of $\{1,2,...,n\}$ of cardinal k, which is independent of V and V' (we assume the $\binom{n}{k}$ subsets of cardinal k have equal probability to be drawn).

For $j \in \{1,2,...,n\}$, define U_j as follows:

[A.19] We do not provide a formal proof of this claim. The interested reader can check its validity by going through the proof of inference results in Chernozhukov et al. (2018) and by verifying that the proof holds when the final p-value is defined as the median of the p-values associated with each sample, multiplied by the ratio between the upper bound for $\mathbb{P}(M_n \leq \alpha)$ and α.

$$U_j = \begin{cases} \alpha V' & \text{if } V \leq 2\alpha \text{ and } j \in A, \\ \alpha + (1-\alpha)V' & \text{otherwise.} \end{cases}$$

It is straightforward to show that the random variables U_j are uniformly distributed and that their lower median, M_n, verifies $\mathbb{P}(M_n \leq \alpha) = \mathbb{P}(V \leq 2\alpha) = 2\alpha$.

Similarly, when n is an odd number, it is possible to construct a sequence of $n = 2k+1$ uniform random variables such that $\mathbb{P}(M_{2k+1} \leq \alpha) = \frac{2k+1}{k+1}\alpha$, and the bound converges to 2α when $n \to \infty$.

Implications. The previous example demonstrates that one cannot improve on the bound proposed by Chernozhukov et al. (2018). Nevertheless, it suggests that this bound is only attained under specific mathematical constructs. In the empirical application we are interested in, the uniform variables are the p-values associated with a test statistic that is estimated on a series of estimation samples under the null hypothesis that the parameter is equal to 0. Such a sequence of p-values has virtually no chance of matching the example above.

Instead, we believe that the polar cases 1 and 2 provide useful insights about the true p-value of the test statistics we are interested in. In our application, the estimation samples on which the statistics are estimated will usually contain common observations. The number of splits of the initial sample is also finite, which implies that iterating the splitting many times will eventually lead to draw estimation samples very similar to previously drawn samples. Therefore, the p-values that are recovered after each estimation are not drawn from independent uniform distributions and, when the number of data splits becomes large enough, each additional p-value will be perfectly correlated with a previous one.

On the other hand, the p-values obtained on two distinct subsamples are unlikely to be perfectly correlated. Consider, for example, the situation in which the whole sample is split into two subsamples, S_1 and S_2, that are used alternatively as training and estimation samples. In this case, it is not clear that there is any dependence between the p-values obtained from S_1 and S_2 used as estimation samples after training a ML predictor on the other sample. More generally, it seems that there is "some independence" between the p-values obtained on different sample splits, and that this independence can help to increase the precision of the final estimate.[A.20]

In the polar case with a perfect positive correlation between the p-values obtained on different subsamples, the p-value associated with the final estimate that we report (which is the median of estimates obtained on different subsamples) is the median of the p-values. In the polar case with no correlation between the p-values, the final p-value is even lower and converges to 0. Based on these observations and the informal discussion above, we argue that the median of the p-values is likely to provide a more realistic approximation of the final estimate's true p-value than twice this median.

[A.20] To put this argument into context, recall that the procedure of Chernozhukov et al. (2018) requires splitting the initial sample into two halves. This random splitting implies that estimates obtained using the estimation sample are less precise than those that could be obtained using the whole sample (standard errors are multiplied by approximately $\sqrt{2}$). The informal arguments put forward in this section suggest that using several random splits instead of one could help limit this loss in precision.

Table M31 – Performance of Alternative Machine Learning Methods in Predicting Heterogeneity in Treatment Effect on Student Outcomes for Girls in Grade 12

	Machine learning method				
	Elastic Net (1)	Random Forest (2)	Linear Model (3)	Boosting (4)	Neural Network (5)
Panel A. Best ML method for the Best Linear Predictor (BLP) of the CATE $s_0(Z)$					
Enrollment outcomes					
Undergraduate major: selective STEM	**0.042238**	0.021458	0.031045	0.021439	0.032774
Undergraduate major: male-dominated STEM	0.015215	0.009745	**0.018221**	0.016650	0.013714
Student perceptions					
Positive perceptions of science-related careers (index)	**0.059065**	0.045943	0.036734	0.030315	0.042442
More men in science-related jobs	**0.030392**	0.015389	0.013050	0.015633	0.014397
Equal gender aptitude for math (index)	0.049497	**0.083063**	0.062788	0.032800	0.065229
Women like science less than men	0.010640	0.011121	**0.014735**	0.011117	0.012712
Women face discrimination in science-related jobs	0.026151	**0.072647**	0.054333	0.019557	0.055785
Taste for science subjects (index)	0.082039	0.074448	**0.090276**	0.050184	0.056377
Math self-concept (index)	0.058826	0.088227	**0.111541**	0.048768	0.066648
Science-related career aspirations (index)	0.072161	0.083165	**0.125844**	0.041561	0.114183
Panel B. Best ML method for the Sorted Group Average Treatment Effects (GATEs)					
Enrollment outcomes					
Undergraduate major: selective STEM	**0.004658**	0.002877	0.003493	0.003283	0.003466
Undergraduate major: male-dominated STEM	**0.003017**	0.002973	0.003003	0.002924	0.002867
Student perceptions					
Positive perceptions of science-related careers (index)	**0.106512**	0.103018	0.101303	0.096945	0.103517
More men in science-related jobs	**0.016919**	0.016358	0.016330	0.016340	0.015974
Equal gender aptitude for math (index)	0026251	**0.029338**	0.027608	0.023143	0.027779
Women like science less than men	0.003346	0.002804	**0.003524**	0.003113	0.003240
Women face discrimination in science-related jobs	0.015339	**0.018028**	0.017264	0.013848	0.016518
Taste for science subjects (index)	0.018653	0.017948	**0.020081**	0.013736	0.016316
Math self-concept (index)	0.012094	0.016213	**0.019361**	0.010633	0.013115
Science-related career aspirations (index)	0.022857	0.024756	**0.034733**	0.019740	0.030833

Notes: This table compares the performance of alternative machine learning (ML) methods in predicting heterogeneity in the treatment effects of the program on student outcomes for girls in Grade 12 (science track), using the approach developed by Chernozhukov et al. (2018). For each outcome, the conditional average treatment effect (CATE) of role model interventions, $s_0(Z)$, is predicted using five alternative ML methods: Elastic Net, Random Forest, Linear Model, Boosting, and Neural Network. The covariates Z that are used to predict the CATE consist of three indicators for the educational districts of Paris, Créteil, and Versailles, four indicators for students' socioeconomic background (high, medium-high, medium-low, and low), their age, their overall percentile rank in the *Baccalauréat* exam, their percentile ranks in the French and math tests of the exam, and a vector of 56 role model fixed effects. For each outcome, Panel A compares the performance of the five ML methods based on the Best Linear Predictor (BLP) targeting of the CATE, whereas Panel B compares their performance based on the Sorted Group Average Treatment Effects (GATEs) targeting of the CATE. In Panel A, the performance measure for the ML learning methods is $\hat{\Lambda} \equiv |\hat{\beta}_2|^2 \widehat{\text{Var}}(S(Z))$, where $S(Z)$ is the ML proxy predictor of the CATE and $\hat{\beta}_2$ is the estimated heterogeneity loading (HET) parameter in the best linear predictor of the CATE. In Panel B, the performance measure for the ML learning methods is $\widehat{\bar{\Lambda}} \equiv \frac{1}{K5} \sum_{k=1}^{5} \hat{\gamma}_k^2$, where $\hat{\gamma}_k$ is the estimated GATE for the quintile k induced by the ML proxy predictor $S(Z)$. For each outcome, the best method (highlighted in bold) is chosen as the one maximizing the performance measure.

Table M32 – Heterogeneous Treatment Effect on Student Outcomes for Girls in Grade 12: Estimates Based on Machine Learning Methods

Panel A. Best Linear Predictor (BLP) of the CATE $s_0(Z)$ given the ML proxy $S(Z)$

Parameters:	ATE (β_1)	HET (β_2)	Best ML method
(p-values in square brackets)			
Undergraduate major: selective STEM	0.038	0.762	Elastic Net
	[0.027]	[0.031]	
Undergraduate major: male-dominated STEM	0.036	0.088	Linear model
	[0.064]	[0.731]	
Positive perceptions of science-related careers (index)	0.298	0.400	Elastic Net
	[0.000]	[0.555]	
More men in science-related jobs	0.119	0.657	Elastic Net
	[0.000]	[0.593]	
Equal gender aptitude for math (index)	0.117	0.324	Random Forest
	[0.010]	[0.108]	
Women like science less than men	0.044	0.095	Linear model
	[0.002]	[0.566]	
Women face discrimination in science-related jobs	0.105	0.496	Random Forest
	[0.000]	[0.012]	
Taste for science subjects (index)	0.008	0.170	Linear Model
	[1.000]	[0.137]	
Math self-concept (index)	0.029	0.257	Linear Model
	[0.988]	[0.010]	
Science-related career aspirations (index)	0.077	0.245	Linear Model
	[0.263]	[0.013]	

Panel B. Average predicted treatment effects among the most/least affected groups (GATEs)

Heterogeneity group:	20% least affected	20% most affected	Difference most−least	Best ML method
(p-values in square brackets)				
Undergraduate major: selective STEM	−0.004	0.139	0.149	Elastic Net
	[1.000]	[0.014]	[0.026]	
Undergraduate major: male-dominated STEM	0.026	0.061	0.038	Elastic Net
	[1.000]	[0.464]	[1.000]	
Positive perceptions of science-related careers (index)	0.316	0.400	0.104	Elastic Net
	[0.037]	[0.001]	[1.000]	
More men in science-related jobs	0.096	0.160	0.065	Elastic Net
	[0.048]	[0.022]	[0.766]	
Equal gender aptitude for math (index)	0.019	0.246	0.210	Random Forest
	[1.000]	[0.037]	[0.332]	
Women like science less than men	0.026	0.073	0.039	Linear model
	[0.758]	[0.078]	[0.772]	
Women face discrimination in science-related jobs	−0.007	0.195	0.197	Random Forest
	[1.000]	[0.003]	[0.038]	
Taste for science subjects (index)	−0.112	0.138	0.251	Linear model
	[0.594]	[0.369]	[0.196]	
Math self-concept (index)	−0.122	0.191	0.317	Linear model
	[0.416]	[0.063]	[0.035]	
Science-related career aspirations (index)	−0.142	0.268	0.387	Linear model
	[0.394]	[0.047]	[0.041]	

Notes: This table reports heterogeneous treatment effects of the program on student outcomes for girls in Grade 12 (science track), using the methods developed by Chernozhukov et al. (2018). For each outcome, the conditional average treatment effect (CATE) of role model interventions, $s_0(Z)$, is predicted using five alternative ML methods: Elastic Net, Random Forest, Linear Model, Boosting, and Neural Network. The covariates Z that are used to predict the CATE consist of three indicators for the educational districts of Paris, Créteil, and Versailles, four indicators for students' socioeconomic background (high, medium-high, medium-low, and low), their age, their overall percentile rank in the *Baccalauréat* exam, their percentile ranks in the French and math tests of the exam, and a vector of 56 role model fixed effects. For each outcome, Panel A reports the parameter estimates and p-values (in square brackets) of the Best Linear Predictor (BLP) of the CATE using the best ML method (see Appendix Table M31, Panel A). The coefficients β_1 and β_2 correspond to the average treatment effect (ATE) and heterogeneity loading (HET) parameters in the BLP, respectively. Panel B reports the Sorted Group Average Treatment Effects (GATEs), i.e., the average treatment effects among students in the top and bottom quintiles of the heterogeneous effects induced by the ML proxy predictor $S(Z)$, using the best ML method (see Appendix Table M31, Panel B). The parameter estimates and p-values are computed as medians over 100 splits, with nominal levels adjusted to account for the splitting uncertainty. This adjustment implies that the reported p-values should be interpreted as upper bounds for the actual p-values.

Table M33 – Heterogeneous Treatment Effects on Selective and Male-Dominated STEM Enrollment for Boys in Grade 12: Estimates based on Machine Learning Methods

Panel A. Best Linear Predictor (BLP) of the CATE $s_0(Z)$ given the ML proxy $S(Z)$			
Parameters:	ATE (β_1)	HET (β_2)	Best ML method
Undergraduate Major: selective STEM	0.005	0.211	Linear Model
p-value	[1.000]	[0.029]	
Undergraduate Major: male-dominated STEM	0.015	0.090	Linear Model
p-value	[1.000]	[0.706]	

Panel B. Sorted Group Average Treatment Effects (GATEs): 20% most and least affected students				
Heterogeneity group:	20% least affected	20% most affected	Difference most−least	Best ML method
Undergraduate Major: selective STEM	−0.056	0.061	0.116	Linear Model
p-value	[0.358]	[0.283]	[0.086]	
Undergraduate Major: male-dominated STEM	0.051	0.010	−0.030	Boosting
p-value	[0.771]	[1.000]	[1.000]	

Panel C. Average characteristics of the 20% most and least affected students (CLAN)				
Heterogeneity group:	20% least affected	20% most affected	Difference most−least	p-value (upper bound)
Enrollment in selective STEM major				
Student characteristics				
Baccalauréat percentile rank in math	48.64	53.26	4.03	0.194
Baccalauréat percentile rank in French	39.95	50.94	10.45	0.000
High SES	0.495	0.494	−0.004	1.000
Role model characteristics				
Professional	0.395	0.600	0.214	0.000
Participated in the program the year before	0.200	0.275	0.070	0.112
Non-French	0.141	0.188	0.051	0.208
Has children	0.413	0.492	0.080	0.140
Age	32.08	33.73	1.58	0.001
Holds/prepares for a Ph.D.	0.707	0.664	−0.070	0.206
Field: Math, Physics, Engineering	0.359	0.236	−0.133	0.001
Field: Earth and Life Sciences	0.541	0.688	0.157	0.000
Enrollment in male-dominated major				
Student characteristics				
Baccalauréat percentile rank in math	54.72	50.21	−4.46	0.123
Baccalauréat percentile rank in French	45.41	47.25	1.38	1.000
High SES	0.465	0.527	0.068	0.248
Role model characteristics				
Professional	0.484	0.531	0.052	0.436
Participated in the program the year before	0.191	0.172	−0.019	1.000
Non-French	0.154	0.124	−0.025	0.820
Has children	0.489	0.489	0.004	1.000
Age	33.32	34.34	0.16	1.000
Holds/prepares for a Ph.D.	0.660	0.682	0.020	1.000
Field: Math, Physics, Engineering	0.295	0.277	−0.015	1.000
Field: Earth and Life Sciences	0.576	0.654	0.074	0.167

Notes: This table reports heterogeneous treatment effects of the program on the undergraduate enrollment outcomes of boys in Grade 12 (science track), using the methods developed by Chernozhukov et al. (2018). For each outcome, the conditional average treatment effect (CATE) of role model interventions, $s_0(Z)$, is predicted using five alternative ML methods: Elastic Net, Random Forest, Linear Model, Boosting, and Neural Network. The covariates Z that are used to predict the CATE consist of three indicators for the educational districts of Paris, Créteil, and Versailles, four indicators for students' socioeconomic background (high, medium-high, medium-low, and low), their age, their overall percentile rank in the *Baccalauréat* exam, their percentile ranks in the French and math tests of the exam, and a vector of 56 role model fixed effects. For each outcome, Panel A reports the parameter estimates and p-values (in square brackets) of the Best Linear Predictor (BLP) of the CATE using the best ML method (see Appendix Table M31, Panel A). The coefficients β_1 and β_2 correspond to the average treatment effect (ATE) and heterogeneity loading (HET) parameters in the BLP, respectively. Panel B reports the Sorted Group Average Treatment Effects (GATEs), i.e., the average treatment effects among students in the top and bottom quintiles of the heterogeneous effects induced by the ML proxy predictor $S(Z)$, using the best ML method (see Appendix Table M31, Panel B). Panel C performs a Classification Analysis (CLAN) by comparing the average characteristics of the 20 percent most and least affected students defined in terms of the ML proxy predictor. The parameter estimates and p-values are computed as medians over 100 splits, with nominal levels adjusted to account for the splitting uncertainty. This adjustment implies that the reported p-values should be interpreted as upper bounds for the actual p-values. Further details on the methods are provided in Appendix M.

Table M34 – Proportion of Random Data Splits for which the Correlation between Conditional Average Treatment Effects (CATEs) can be Computed, Girls in Grade 12

	Proportion of data splits such that			
	$\hat{\rho}_{A,B\|Z}$ can be computed*	$\hat{\beta}_2^{B\|B} > 0$	$\hat{\beta}_2^{A\|A} > 0$	$\hat{\beta}_2^{A\|B}\hat{\beta}_2^{B\|A} \geq 0$
	(1)	(2)	(3)	(4)
When outcome Y^B is enrollment in a selective STEM program and outcome Y^A is:				
Positive perception of science-related careers (index)	0.80	1.00	0.86	0.90
More men in science-related jobs	0.68	0.99	0.89	0.73
Equal gender aptitude for math (index)	0.35	1.00	0.98	0.36
Women like science less than men	0.34	0.99	0.84	0.40
Women face discrimination in science-related jobs	0.62	1.00	1.00	0.62
Taste for science subjects (index)	0.81	0.99	0.97	0.83
Math self-concept (index)	0.39	0.99	1.00	0.40
Science-related career aspirations (index)	0.64	0.99	1.00	0.65
Number of data splits	3,000	3,000	3,000	3,000

Notes: This table reports, for the sample of girls in Grade 12 (science track), the proportion of random data splits (out of 3,000) for which the correlation between the Conditional Average Treatment Effects (CATEs) on outcomes Y^A and Y^B could be computed. Outcome Y^B is always enrollment in selective STEM while Y^A is the outcome listed in the corresponding row of the table. Conditional on the covariates Z, the CATEs on outcomes Y^A and Y^B are denoted by $s_0^A(Z)$ and $s_0^B(Z)$, respectively, whereas their ML proxy predictors are denoted by $S^A(Z)$ and $S^B(Z)$, respectively. For each random split, the correlation coefficient $\rho_{A,B|Z}$ is estimated as $\hat{\rho}_{A,B|Z} = \text{Sign}(\hat{\beta}_2^{A|B})\sqrt{\hat{\beta}_2^{A|B}\hat{\beta}_2^{B|A}}/\sqrt{\hat{\beta}_2^{A|A}}\sqrt{\hat{\beta}_2^{B|B}}$, where $\hat{\beta}_2^{k|l}$ is the estimated heterogeneity loading parameter of the Best Linear Predictor (BLP) of $s_0^k(Z)$ based on $S^l(Z)$ (with $k, l \in \{A, B\}$), using the methods in Chernozhukov et al. (2018). Column 1 indicates the fraction of data splits for which $\hat{\rho}_{A,B|Z}$ could be computed. The next three columns report the fraction of sample splits for which each of the three conditions to compute $\hat{\rho}_{A,B|Z}$ is met, i.e., $\hat{\beta}_2^{B|B} > 0$ (column 2), $\hat{\beta}_2^{A|A} > 0$ (column 3), and $\hat{\beta}_2^{A|B}\hat{\beta}_2^{B|A} \geq 0$ (column 4). The proportion of random splits such that $\beta_2^{B|B} > 0$ varies slightly across rows because for each pair of outcomes (Y^A,Y^B), the sample is restricted to observations with non-missing values for both outcomes (see Appendix M). Table 9 in the main text reports the median and 95 percent confidence interval of $\hat{\rho}_{A,B|Z}$ over the first 100 random data splits for which $\hat{\rho}_{A,B|Z}$ can be computed. Details are provided in Section 6.4 of the main text and in Appendix M.

Table M35 – Correlation between Conditional Average Treatment Effects (CATEs) for Girls in Grade 12: Sensitivity Analysis

	Bivariate correlation with the CATE on enrollment in a selective STEM program (from first 100 valid iterations)		
	Estimate ($\hat{\rho}_{A,B\|Z}$)	95% confidence interval	Proportion of valid iterations
Panel A. Data splits such that $\hat{\beta}_2^{A\|A} > 0.1$, $\hat{\beta}_2^{B\|B} > 0.1$ and $\hat{\beta}_2^{A\|B}\hat{\beta}_2^{B\|A} \geq 0$			
Conditional average treatment effect (CATE) on:			
Positive perception of science-related careers (index)	0.96	[0.21, 5.39]	0.73
More men in science-related jobs	−0.68	[−3.33, −0.03]	0.65
Equal gender aptitude for math (index)	0.08	[−1.90, 2.11]	0.33
Women like science less than men	0.26	[−0.64, 3.75]	0.19
Women face discrimination in science-related jobs	−0.31	[−2.20, 0.61]	0.61
Taste for science subjects (index)	0.69	[0.07, 3.42]	0.66
Math self-concept (index)	−0.06	[−1.85, 1.37]	0.38
Science-related career aspirations (index)	0.34	[−0.61, 1.95]	0.62
Panel B. Data splits such that $\hat{\beta}_2^{A\|A} > 0.2$, $\hat{\beta}_2^{B\|B} > 0.2$ and $\hat{\beta}_2^{A\|B}\hat{\beta}_2^{B\|A} \geq 0$			
Conditional average treatment effect (CATE) on:			
Positive perception of science-related careers (index)	0.93	[0.21, 5.07]	0.64
More men in science-related jobs	−0.68	[−3.26, −0.03]	0.65
Equal gender aptitude for math (index)	0.05	[−1.98, 1.90]	0.31
Women like science less than men	0.31	[−0.51, 3.44]	0.05
Women face discrimination in science-related jobs	−0.30	[−2.12, 0.64]	0.58
Taste for science subjects (index)	0.59	[0.07, 2.61]	0.34
Math self-concept (index)	0.05	[−1.68, 1.51]	0.29
Science-related career aspirations (index)	0.31	[−0.64, 1.79]	0.46

Notes: Similarly to Table 9 in the main text, this table reports, for girls in Grade 12 (science track), the estimates of the bivariate correlation $\rho_{A,B|Z}$ between the Conditional Average Treatment Effect (CATE) on enrollment in a selective STEM program, denoted by $s_0^B(Z)$, and the CATE on each of the potential mediators listed in the table, denoted by $s_0^A(Z)$. The difference is that estimates provided in this table are obtained using only iterations of the data-splitting process for which the estimates of the heterogeneity loading parameters $\hat{\beta}_2^{A|A}$ and $\hat{\beta}_2^{B|B}$ are above a certain threshold. This threshold is set at 0.1 in Panel A and at 0.2 in Panel B. These restrictions are applied to check the sensitivity of the correlation estimates to excluding data splits that yield a poor ML prediction of the CATEs on outcomes Y^A or Y^B. Column 3 indicates the proportion of data splits satisfying the restrictions specified in each panel's heading. The estimates and 95 percent confidence intervals reported in columns 1 and 2 are obtained using the first 100 data splits satisfying these restrictions. Additional details are provided in the notes of Table 9 and in Appendix M.

Appendix References

Anderson, Michael L., "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*, 2008, *103* (484).

Athey, Susan and Guido W. Imbens, "The Econometrics of Randomized Experiments," in Esther Duflo and Abhijit V. Banerjee, eds., *Handbook of Economic Field Experiments*, Vol. 1, Elsevier, 2017, pp. 73–140.

Baron, Reuben M. and David A. Kenny, "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 1986, *51* (6), 1173–1182.

Beede, David, Tiffany Julian, David Langdon, George McKittrick, Beethika Khan, and Mark Doms, "Women in STEM: A Gender Gap to Innovation," 2011. U.S. Department of Commerce, Economics and Statistics Administration, Issue Brief No. 04-11.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli, "Adaptive Linear Step-up Procedures that Control the False Discovery Rate," *Biometrika*, 2006, *93* (3), 491–507.

Bloom, Erik, Indu Bhushan, David Clingingsmith, Rathavuth Hong, Elizabeth King, Michael Kremer, Benjamin Loevinsohn, and J. Brad Schwartz, "Contracting for Health: Evidence from Cambodia," 2006. Brookings Institution.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," 2018. NBER Working Paper No. 24678.

Cohen, Jessica and Pascaline Dupas, "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics*, 2010, *125* (1), 1–45.

Davison, Anthony C. and David V. Hinkley, *Bootstrap Methods and their Application*, Cambridge University Press, 1997.

Duflo, Esther and Emmanuel Saez, "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," *The Quarterly Journal of Economics*, 2003, *118* (3), 815–842.

Dupas, Pascaline, Elise Huillery, and Juliette Seban, "Risk Information, Risk salience, and Adolescent Sexual Behavior: Experimental Evidence from Cameroon," *Journal of Economic Behavior & Organization*, 2018, *145*, 151–175.

Fisher, Ronald A., *The Design of Experiments*, McMillan, 1935.

Frölich, Markus and Martin Huber, "Direct and Indirect Treatment Effects–Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017, *79* (5), 1645–1666.

Fujiwara, Thomas and Leonard Wantchekon, "Can Informed Public Deliberation Overcome Clientelism? Experimental Evidence from Benin," *American Economic Journal: Applied Economics*, 2013, *5* (4), 241–255.

Heckman, James, Rodrigo Pinto, and Peter Savelyev, "Understanding the Mechanisms through which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 2013, *103* (6), 2052–2086.

Ichino, Nahomi and Matthias Schündeln, "Deterring or Displacing Electoral Irregularities? Spillover Effects of Observers in a Randomized Field Experiment in Ghana," *The Journal of Politics*, 2012, *74* (1), 297–307.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto, "Experimental Designs for Identifying Causal Mechanisms," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2013, *176* (1), 5–51.

_ **, Luke Keele, and Teppei Yamamoto**, "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 2010, pp. 51–71.

Imbens, Guido W. and Donald B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, 2015.

Keele, Luke, "Causal Mediation Analysis: Warning! Assumptions Ahead," *American Journal of Evaluation*, 2015, *36* (4), 500–513.

Kuhn, Max, "Building Predictive Models in R using the caret Package," *Journal of Statistical Software*, 2008, *28* (5), 1–26.

MacKinnon, David P., Amanda J. Fairchild, and Matthew S. Fritz, "Mediation Analysis," *Annual Review of Psychology*, 2007, *58*, 593–614.

McDonald, Judith A. and Robert J. Thornton, "Do New Male and Female College Graduates Receive Unequal Pay?," *Journal of Human Resources*, 2007, *42* (1), 32–48.

Paz, Lourenço S. and James E. West, "Should We Trust Clustered Standard Errors? A Comparison with Randomization-Based Methods," 2019. NBER Working Paper No. 25926.

Rosenbaum, Paul R., *Observational Studies*, Springer, 2002.

_ , *Design of Observational Studies*, Springer Series in Statistics, 2010.

Vazquez-Bare, Gonzalo, "Identification and Estimation of Spillover Effects in Randomized Experiments," 2018. Manuscript.

Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.