(For Online Publication)

Appendix to

How Effective are Female Role Models in Steering Girls towards STEM? Evidence from French High Schools

Thomas Breda
Julien Grenet
Marion Monnet
Clémentine Van Effenterre

The Economic Journal

January 2023

List of Appendices

A 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 analyses 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 study can be expected to reduce the gender pay gap. Section A.1 describes the data sources, while Section A.2 discusses the empirical results.

A.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 is based on the combination of aggregate statistics on student enrolment 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 analysed 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 (MESRI-DGESIP/DGRI-SIES, 2017) and provides detailed information on each student's degree program and field of study.

Enquête d'Insertion Professionnelle à 30 Mois des Diplômés de Master 2015 (EIPDM). This survey was conducted in December 2017 by the Ministry of Higher Education (MESRI, 2018) to collect information on the transition of master's graduates to the labour market. The survey was targeted at students who obtained their master's degree in 2015 and who entered the labour market within one year after graduation, with an overall response rate of 70%. 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, 2018), 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 were members of the CGE in 2018, with an overall response rate of 48%. 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.

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

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 *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 (*grandes 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%. This distinction does not apply to selective STEM majors, since almost all engineering schools are male-dominated.

Earnings information. The EIPDM 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.

A.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 analyse 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 A1 show the distribution of master's-level graduates across the broad categories of college majors defined above, along with the share of female graduates in each category. The summary statistics indicate that while female students represent 52% of master's level graduates, they are strongly under-represented in STEM majors (34%). Female under-representation is more pronounced in selective (male-dominated) STEM majors (female share: 30%) than in non-selective STEM majors (female share: 40%). Among non-selective STEM majors, female students represent only 29% of graduates in male-dominated fields such as mathematics, physics or computer science, compared to 60% 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 over-represented in lower-paying majors (see columns 3–5 of Table A1). Female graduates holding a STEM degree have a median starting wage of €29,984, which is 7.4% higher than the median starting wage of female graduates holding a non-STEM degree (€27,913). Strikingly, the wage premium for female graduates in STEM appears to be almost entirely driven by selective (male-dominated) STEM degrees (16.4%). By contrast, the wage premium attached to non-selective STEM degrees is close to zero (−0.5%). 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% for female graduates compared to non-STEM majors, a wage penalty of 4.7% is attached to female-dominated non-selective STEM majors.

Female under-representation in STEM: contribution to the gender pay gap. The last three columns of Table A1 indicate that across all categories of programs, male graduates earn a median annual starting wage of €32,122, compared to €28,411 for female graduates. This amounts to an overall gender pay gap of €3,711 per year, or 11.6% of male pay.

Although the over-representation 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 under-representation in STEM majors (STEM versus non-STEM, selective versus non-selective STEM, male-dominated versus 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 versus 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 versus 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 under-representation 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 similarly, 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 A2 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% of the gender pay gap among college graduates. Two-thirds of this explained part (27.7% of the total wage gap) can be attributed to the unequal representation of female and male graduates in STEM versus 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 (12.3% of the total) is due to gender segregation between non-STEM majors, the lowest-paying majors (humanities) being typically more female-dominated (77%) than the highest-paying ones (law and economics, where the female share is 59%).

The 27.7% 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% reduction in the gender pay gap. In line with the evidence from Table A1, 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% 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% reduction in the gender pay gap, while further reassigning female students between majors within male- and female-dominated programs would result in an extra 2.9% reduction from the baseline.

Altogether, these findings suggest that the under-representation of female students in STEM majors accounts for approximately 28% of the entry-level gender pay gap among college graduates in France. 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 A1 – 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 (15)	18,874	11.3	29.4	29,077	104.2	31,371	100.2	7.3
Female-dominated majors (3)	9,266	5.6	60.3	26,596	95.3	27,581	88.1	3.6

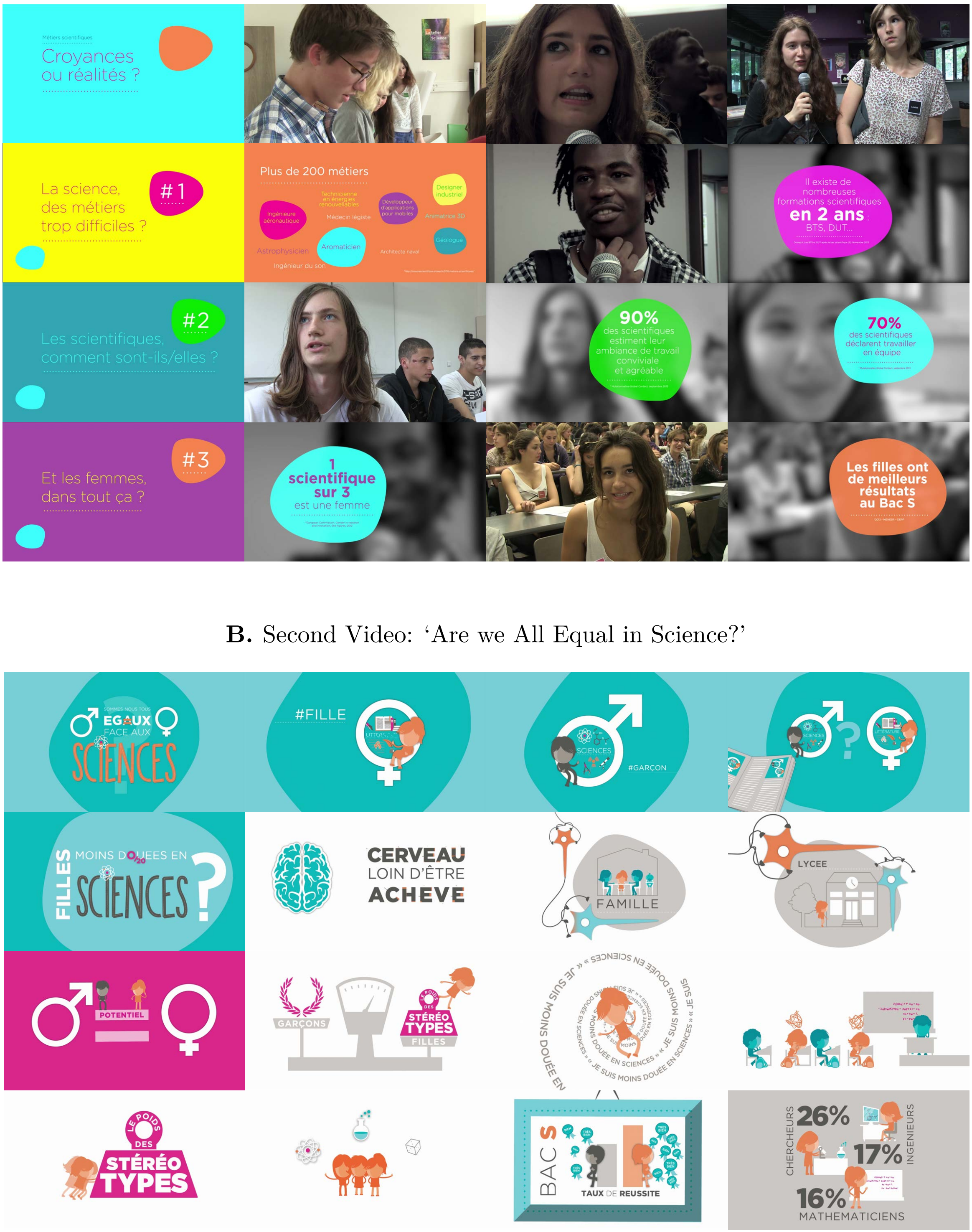
Notes: This table reports summary statistics on gender segregation and gender pay gaps for the population of college graduates who obtained their master's degree (or equivalent) in 2015 or 2016. The 54 college majors are grouped into two broad categories: non-STEM majors (master's degrees in economics, 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 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%. 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 EIPDM 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) (CGE, 2018); columns 4–8: Enquête d'Insertion Professionnelle à 30 Mois des Diplômés de Master 2015 (EIPDM) (MESRI, 2018) and EIDGE.

Table A2 – 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	0.032	27.7%
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	0.003	2.9%
between majors within non-STEM	0.014	12.3%
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 major, 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 similarly, 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 A1.

B Program Details



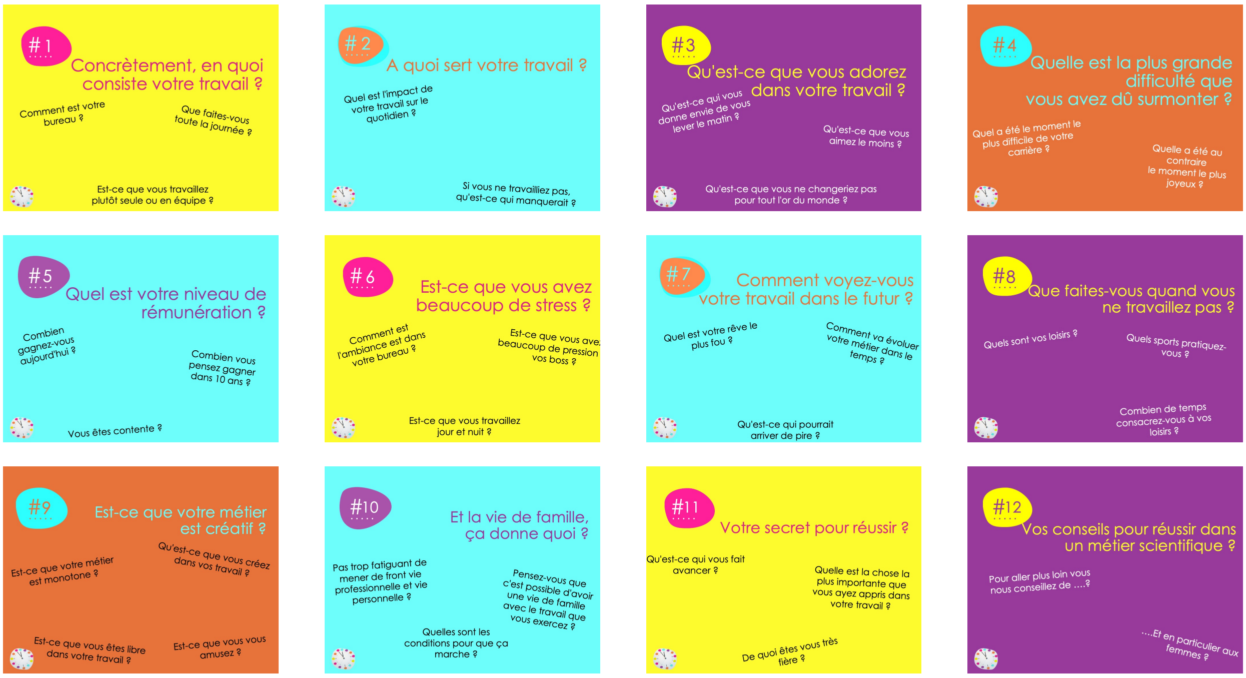
A. First Video: ‘Jobs in Science: Beliefs or Reality?’

Métiers scientifiques
Croyances ou réalités ?
La science, des métiers trop difficiles ?
#1
Plus de 200 métiers
Technicienne en énergies renouvelables
Designer industriel
Ingénieure aéronautique
Développeur d'applications pour mobiles
Médecin légiste
Animatrice 3D
Astrophysicien
Aromaticien
Architecte naval
Géologue
Ingénieur du son
Il existe de nombreuses formations scientifiques en 2 ans : BTS, DUT...
Les scientifiques, comment sont-ils/elles ?
#2
90% des scientifiques estiment leur ambiance de travail conviviale et agréable
70% des scientifiques déclarent travailler en équipe
Et les femmes, dans tout ça ?
#3
1 scientifique sur 3 est une femme
Les filles ont de meilleurs résultats au Bac S

B. Second Video: ‘Are we All Equal in Science?’

SOMMES-NOUS TOUS EGAUX FACE AUX SCIENCES
#FILLE
LITTÉRATURE
SCIENCES
#GARÇON
FILLES MOINS DOUÉES EN SCIENCES ?
CERVEAU LOIN D'ÊTRE ACHEVE
FAMILLE
LYCEE
POTENTIEL
GARÇONS
LE POIDS DES STÉRÉO TYPES
FILLES
« JE SUIS MOINS DOUÉE EN SCIENCES »
BAC S
TAUX DE REUSSITE
CHERCHEURS
26%
17%
INGENIEURS
16%
MATHEMATICIENS

Figure B1 – Screenshots of the Two Videos Shown During the Role Model Interventions



#1

Concrètement, en quoi consiste votre travail ?

Comment est votre bureau ?

Que faites-vous toute la journée ?

Est-ce que vous travaillez plutôt seule ou en équipe ?

#2

A quoi sert votre travail ?

Quel est l'impact de votre travail sur le quotidien ?

Si vous ne travailliez pas, qu'est-ce qui manquerait ?

#3

Qu'est-ce que vous adorez dans votre travail ?

Qu'est-ce qui vous donne envie de vous lever le matin ?

Qu'est-ce que vous aimez le moins ?

Qu'est-ce que vous ne changeriez pas pour tout l'or du monde ?

#4

Quelle est la plus grande difficulté que vous avez dû surmonter ?

Quel a été le moment le plus difficile de votre carrière ?

Quelle a été au contraire le moment le plus joyeux ?

#5

Quel est votre niveau de rémunération ?

Combien gagnez-vous aujourd'hui ?

Combien vous pensez gagner dans 10 ans ?

Vous êtes contente ?

#6

Est-ce que vous avez beaucoup de stress ?

Comment est l'ambiance est dans votre bureau ?

Est-ce que vous ave beaucoup de pression vos boss ?

Est-ce que vous travaillez jour et nuit ?

#7

Comment voyez-vous votre travail dans le futur ?

Quel est votre rêve le plus fou ?

Comment va évoluer votre métier dans le temps ?

Qu'est-ce qui pourrait arriver de pire ?

#8

Que faites-vous quand vous ne travaillez pas ?

Quels sont vos loisirs ?

Quels sports pratiquez-vous ?

Combien de temps consacrez-vous à vos loisirs ?

#9

Est-ce que votre métier est créatif ?

Est-ce que votre métier est monotone ?

Qu'est-ce que vous créez dans vos travail ?

Est-ce que vous êtes libre dans votre travail ?

Est-ce que vous vous amusez ?

#10

Et la vie de famille, ça donne quoi ?

Pas trop fatiguant de mener de front vie professionnelle et vie personnelle ?

Pensez-vous que c'est possible d'avoir une vie de famille avec le travail que vous exercez ?

Quelles sont les conditions pour que ça marche ?

#11

Votre secret pour réussir ?

Qu'est-ce qui vous fait avancer ?

Quelle est la chose la plus importante que vous ayez appris dans votre travail ?

De quoi êtes vous très fière ?

#12

Vos conseils pour réussir dans un métier scientifique ?

Pour aller plus loin vous nous conseillez de ….?

….Et en particulier aux femmes ?

Figure B2 – Screenshots of the Slides Provided to the Role Models to Describe their own Experience

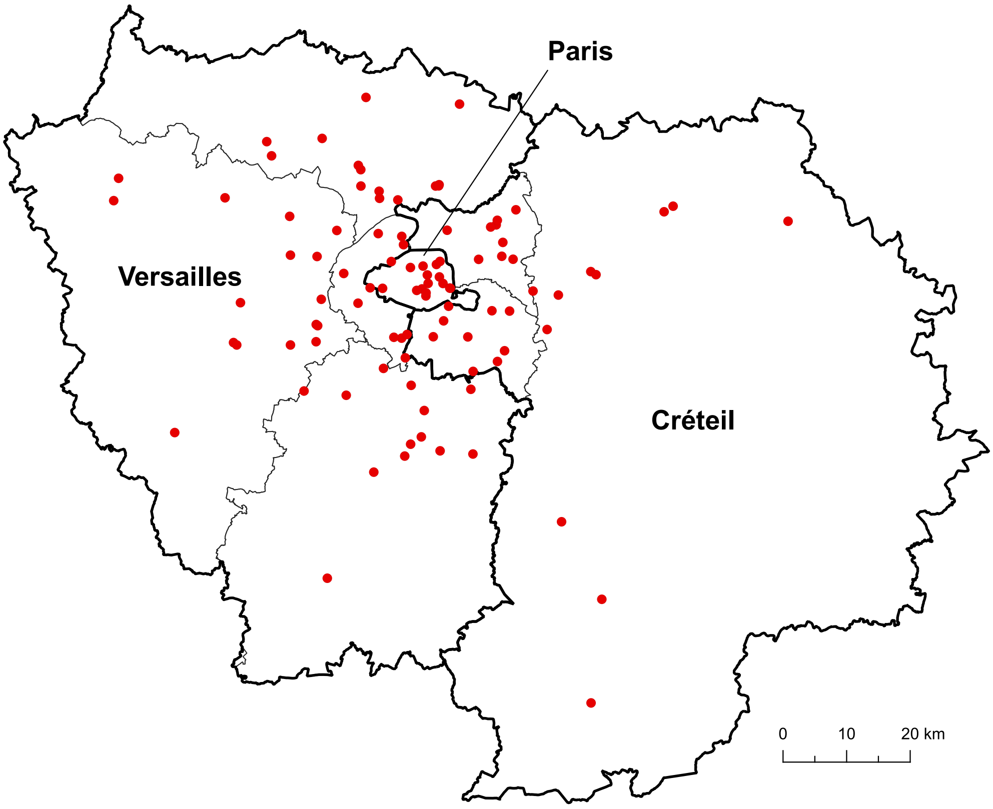


Figure B3 – Participating High Schools

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. The shapefile for the administrative divisions of the Paris region (*départements*) were obtained from APUR (2018).

C Student-Level Administrative Data

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

C.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 enrolment data. Students' socio-demographic characteristics and enrolment status are obtained from the *Bases Élèves Académiques* (BEA) for academic years 2012/13 to 2016/17 (DAPEP, 2017; PAPP, 2017; SSA, 2017). 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. They also cover 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' two-digit occupation and detailed information on their enrolment 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 enrolment data. To track grade 12 (science track) students' enrolment 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) (MESRI-DGESIP/DGRI-SIES, 2017), which is managed by the Statistical Office of the French Ministry of Higher Education (Sous-Direction des Systèmes d'Information et des Études Statistiques). This dataset, which covers the academic years 2012/13 to 2016/17, records all students enrolled in the French higher education system outside 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) (MENJ-DEPP, 2017), contains students' individual exam results for the *diplôme national du brevet* (DNB), which middle school students take at the end of 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).

C.2 Classification of STEM Undergraduate Programs

The enrolment 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 versus non-selective STEM programs.

- *Selective STEM*: This category includes all CPGE programs with a specialisation 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 specialise in STEM fields.
- *Non-selective STEM*: This category includes non-selective university bachelor's degree programs (*licence*) that specialise in STEM fields: maths, physics, chemistry, earth and life sciences, and computer science. Undergraduate programs in medicine and pharmacy are not included in this category.

Male- versus female-dominated STEM programs.

- *Male-dominated STEM*: STEM programs are classified as being male dominated if the share of female students in the corresponding field is below 50%. This category includes the selective programs (CPGE and STS) and non-selective programs (*licence*) that specialise in mathematics, physics, chemistry, computer science and engineering.
- *Female-dominated STEM*: STEM programs are classified as being female dominated if the share of female students in the corresponding field is above 50%. This category includes both selective (CPGE and STS) and non-selective programs (*licence*) that specialise 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.

Note that selective STEM programs and male-dominated STEM programs are partly overlapping: in 2016/17, 49% of undergraduate students in male-dominated STEM fields were enrolled in selective programs, while 95% of students in selective programs were in male-dominated STEM fields.

D Construction of Synthetic Indices and Multiple Hypothesis Testing

This appendix discusses the construction of the synthetic indices that we use to measure the effects of role model interventions on students' perceptions (Section D.1) and provides further details on the adjustment of p-values to correct for multiple hypothesis testing (Section D.2).

D.1 Construction of Synthetic Indices

The student survey questionnaire aimed at measuring the effects of role model 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)* self-concept in maths and *(v)* science-related career aspirations.

We use the survey items listed below to construct synthetic indices for each of these five dimensions. When responses are measured on a Likert scale, i.e., when respondents specify their level of agreement or disagreement with a statement on a symmetric agree-disagree scale, the item responses are recoded so that higher values correspond to less stereotypical or negative perceptions (see details below). We then take the average of each student's responses to the different questions.[A.3] We checked that the indices yield similar results if item responses are converted to binary variables before taking the average across items. Finally, to facilitate interpretation, we normalise each index to have a mean of zero and a standard deviation of one in the control group.

Below is the list of the individual items that are included in each of the five synthetic indices. Unless otherwise specified, these items use a four-point Likert response scale such that 1=Strongly agree, 2=Agree, 3=Disagree and 4=Strongly disagree. Items marked with a * have been recoded such that a value of 1 means 'Strongly disagree' and 4 means 'Strongly agree'.

1. *Positive perceptions of science-related careers* (5 items): 'Science-related jobs require more years of schooling'; 'Science-related jobs are monotonous'; 'Science-related jobs are rather solitary'; 'Science-related jobs pay higher wages*'; 'It is difficult to have a fulfilling family life when working as a scientist'.
2. *Equal gender aptitude for maths* (2 items): 'Women and men are born with different brains'; 'Men are more gifted than women in mathematics'.
3. *Taste for science subject* (4 items): Enjoys maths (on a scale from 0 'not at all' to 10 'very much'); Enjoys physics and chemistry (on a scale from 0 to 10); Enjoys earth and life sciences (on a scale from 0 to 10); 'I like science in general*'.
4. *Self-concept in maths* (4 items): Self-assessed performance in match (very weak/weak/average/good/very good); 'I feel lost when I try to solve a maths problem'; 'I often worry that I will struggle in maths class'; 'If I make enough effort, I can do well in science subjects'.
5. *Science-related career aspirations* (4 items): 'Some jobs in science are interesting*'; 'I could see myself working in a science-related job later in life*'; Interested in at least one of six STEM job out of a list of ten STEM and non-STEM occupations[A.4] (0/1 variable); 'Career and earnings prospects play an important role in my choice of study' (on a scale from 0 'not at all' to 10 'very much').

[A.3] This procedure is inspired from the KidIQol test used in the psychological literature to measure children's life satisfaction (Gayral-Taminh et al., 2005).

[A.4] 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.

D.2 Multiple Hypothesis Testing

Consistent with the recent applied literature, we systematically 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).

We study nine main outcomes throughout the paper: *(i)* enrolment in a STEM track (for grade 10 students) or STEM major (for grade 12 students); *(ii)* five synthetic indices capturing positive perceptions of science-related careers, equal gender aptitude for maths, taste for science subjects, self-concept in maths and science-related career aspirations (see Section D.1); and *(iii)* three variables capturing different facets of gender role in science that cannot be combined into a single index, which are based on the survey items asking students whether they agree or disagree with the statements 'There are more men than women in science-related jobs', 'Women do not really like science' and 'Women face discrimination in science-related jobs'. These nine outcomes are our primary outcomes of interest and we therefore systematically provide (along with standard p-values) p-values that are adjusted for multiple testing across them (q-values), separately by grade level and gender.

For each of the five synthetic indices described in the previous section, we report separate treatment effect estimates for the individual components of the index and provide standard p-values for the corresponding estimates along with p-values adjusted for multiple testing across the index components, separately by grade level and gender.

As we further split enrolment in STEM into different types of STEM tracks or majors (e.g., selective STEM, non-selective STEM, male-dominated STEM and female-dominated STEM in grade 12), we provide adjusted p-values for multiple testing across these different STEM tracks or majors, separately by grade level and gender. Given the importance of some of these specific STEM majors in our analyses, it could also be justified to consider them jointly with the primary outcomes described above. We have checked that, in practice, this alternative choice has little effect on the reported q-values.

Finally, treatment effects on other outcomes, such as the probabilities of being enrolled in a non-STEM major or of not being enrolled in an education program in the year following the classroom interventions, are also reported in the paper for the sake of completeness and clarity. Since these are not outcomes of direct interest in our study or are complements of other outcomes of interest, we do not consider them in the multiple testing corrections.

E Summary Statistics and Balancing Tests

Table E1 – Experimental Sample: Summary Statistics (School-Level)

	High schools operating in the Paris region (1)	Participating high schools (2)
Number of high schools	489	98
Share private	0.339	0.173
Education district: Paris	0.243	0.153
Education district: Créteil	0.348	0.296
Education district: Versailles	0.409	0.551
Number of students	644	924
Share of female students	0.524	0.526
Share of high SES students	0.423	0.391
Share of medium-high SES students	0.116	0.128
Share of medium-low SES students	0.243	0.239
Share of low SES students	0.218	0.241
Pass rate on baccalauréat exam in 2015	0.913	0.910

Notes: This table compares the characteristics of high schools that participated in the program evaluation in 2015/16 to the characteristics of all general-track high schools operating in the Paris region. 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. The *baccalauréat* pass rate is computed for students who were enrolled in grade 12 in 2014/15, i.e., in the year before the intervention, and who took the exams in the general or technical tracks.

Table E2 – Experimental Sample: Summary Statistics (Student-Level)

	High schools operating in the Paris region	Participating high schools			
		Classes selected for random assignment	Classes not selected for random assignment	Diff. (2)−(3)	*p*-value of diff. (2)−(3)
	(1)	(2)	(3)	(4)	(5)
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	0.004	0.503
Age (years)	15.14	15.13	15.14	−0.016	0.004
Non-French	0.063	0.060	0.068	−0.008	0.005
High SES	0.403	0.381	0.361	0.020	0.000
Medium-high SES	0.118	0.128	0.127	0.001	0.713
Medium-low SES	0.239	0.241	0.248	−0.006	0.203
Low SES	0.240	0.249	0.265	−0.015	0.002
Number of siblings	1.44	1.49	1.50	−0.016	0.255
Class size	32.22	33.25	32.48	0.753	0.000
DNB percentile rank in maths	57.69	58.48	55.10	3.382	0.000
DNB percentile rank in French	57.23	57.85	55.75	2.096	0.000
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	0.075	0.000
Age (years)	17.11	17.12	17.10	0.023	0.043
Non-French	0.045	0.051	0.037	0.014	0.000
High SES	0.527	0.464	0.535	−0.071	0.000
Medium-high SES	0.115	0.136	0.126	0.010	0.113
Medium-low SES	0.198	0.209	0.180	0.029	0.000
Low SES	0.160	0.192	0.160	0.032	0.000
Number of siblings	1.43	1.50	1.44	0.054	0.007
Class size	31.43	31.97	32.08	−0.153	0.069
DNB percentile rank in maths	76.25	74.06	76.20	−2.127	0.000
DNB percentile rank in French	70.78	69.61	69.78	−0.169	0.704

Notes: This table compares the characteristics of grade 10 and grade 12 (science track) students enrolled in the high schools that participated in the program evaluation to the characteristics of all grade 10 and grade 12 (science track) students enrolled in general-track high schools in the Paris region. In participating schools, the 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 maths scores are from the exams of the *diplôme national du brevet* (DNB) that middle school students take at the end of grade 9.

Table E3 – 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]	0.600	0.596	0.603	0.007	0.922
Teacher's gender: female	0.551	0.533	0.565	0.032	0.653
Teacher showed interest	0.696	0.634	0.745	0.111	0.098
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	0.169	0.197	0.147	−0.050	0.353
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. Topics addressed during the intervention					
'Science is everywhere'	1.000	1.000	1.000	0.000	–
'Jobs in science are fulfilling'	0.990	1.000	0.982	−0.018	0.080
'Jobs in science are for girls too'	1.000	1.000	1.000	0.000	–
'Jobs in science pay well'	0.866	0.890	0.849	−0.040	0.516
Short videos	0.980	0.969	0.988	0.019	0.436
D. 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
E. 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 classroom 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 are mathematics, physics and chemistry, and earth and life sciences.

Table E4 – Compliance with Random Assignment

		Classes assigned to	
	All classes (1)	Control group (2)	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 E5 – 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 (0.012)	0.026
Number of students	6,801	6,899	13,700	
Panel B. Grade 12 (science track)				
Survey response rate	0.909	0.912	0.005 (0.012)	0.693
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 randomisation was stratified by school. Standard errors (in parentheses) are adjusted for clustering at the unit of randomisation (class).

Table E6 – Treatment-Control Balance: Survey Respondents

			Within school	
	Control group (1)	Treatment group (2)	Difference T−C (3)	*p*-value of diff. (4)
Panel A. Grade 10				
Student characteristics				
Female	0.538	0.521	−0.014	0.160
Age (years)	15.12	15.11	−0.01	0.248
Non-French	0.057	0.060	0.003	0.528
High SES	0.382	0.389	0.005	0.496
Medium- high SES	0.133	0.127	−0.006	0.248
Medium-low SES	0.245	0.235	−0.009	0.200
Low SES	0.240	0.248	0.010	0.158
Number of siblings	1.483	1.482	−0.001	0.954
Class size	33.23	33.25	0.02	0.837
At least one science elective course	0.394	0.402	0.009	0.693
At least one standard elective course	0.773	0.738	−0.032	0.132
DNB percentile rank in maths	59.09	59.04	−0.18	0.760
DNB percentile rank in French	58.14	58.41	0.08	0.893
Test of joint significance	*F*-statistic: 0.634 (*p*-value: 0.813)			
Predicted track in grade 11				
Grade 11: science track	0.454	0.459	0.004	0.577
Grade 11: science–general track	0.381	0.385	0.003	0.666
Grade 11: science–technical track	0.073	0.074	0.001	0.670
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 maths	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*-statistic: 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 randomisation was stratified by school, and standard errors are adjusted for clustering at the unit of randomisation (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., enrolment 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 E7 – Balancing Test: High Schools Visited by Professionals and Researchers, Grade 10 Students

	High school visited by		Difference (2)−(1)	p-value of diff.
	Researcher (1)	Professional (2)	(3)	(4)
School characteristics				
Education district: Paris	0.165	0.167	0.002	0.958
Education district: Créteil	0.273	0.317	0.044	0.321
Education district: Versailles	0.562	0.516	−0.046	0.348
Private school	0.092	0.224	0.132	0.000
Share of female students in 2015/16	0.523	0.527	0.005	0.627
Pass rate on baccalauréat exam in 2015[a]	0.904	0.916	0.012	0.041
Grade 10 students: science track in grade 11[b]	0.405	0.412	0.006	0.597
Grade 10 students: general science track in grade 11[b]	0.341	0.337	−0.005	0.672
Grade 10 students: technical science track in grade 11[b]	0.064	0.075	0.011	0.135
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 maths	57.80	59.02	1.22	0.380
DNB percentile rank in French	56.77	58.71	1.93	0.120
Predicted track in grade 11				
Grade 11: science track	0.448	0.454	0.006	0.668
Grade 11: science–general track	0.374	0.375	0.002	0.915
Grade 11: science–technical track	0.074	0.079	0.005	0.517
N	6,059	7,641	13,700	

Notes: Each row corresponds to a different linear regression with the dependent variable listed on the left for students enrolled in grade 10 in 2015/16. 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 an indicator that takes the value one if the school was visited by a professional and zero if the school was visited by a researcher, with the p-value reported in column 4. Standard errors are adjusted for clustering at the class level. High school tracks in grade 11 are predicted for each student using the coefficients from a linear regression of the corresponding binary variable (e.g., enrolment in the general science track) on all the school and student characteristics listed in the table. This model is fitted on the sample of students in the control group. [a] The *baccalauréat* pass rate is computed for students who were enrolled in grade 12 in 2014/15, i.e., in the year before the intervention, and who took the exams in the general or technical tracks. [b] The share of students enrolled in the science track in grade 11 is computed for students who were enrolled in grade 10 in 2014/15.

Table E8 – Balancing Test: High Schools Visited by Professionals and Researchers, Grade 12 Students

	High school visited by		Difference (2)−(1)	p-value of diff.
	Researcher	Professional		
	(1)	(2)	(3)	(4)
School characteristics				
Education district: Paris	0.164	0.163	−0.001	0.985
Education district: Créteil	0.223	0.321	0.098	0.138
Education district: Versailles	0.614	0.517	−0.097	0.195
Private school	0.096	0.244	0.148	0.007
Share of female students in 2015/16	0.533	0.543	0.010	0.379
Pass rate on baccalauréat exam in 2015[a]	0.911	0.912	0.002	0.849
Grade 12 (science track) students: STEM major in higher ed.[b]	0.409	0.384	−0.025	0.050
Grade 12 (science track) students: selective STEM in higher ed.[b]	0.191	0.202	0.010	0.484
Grade 12 (science track) students: male-dom. STEM in higher ed.[b]	0.309	0.299	−0.010	0.431
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 maths	72.96	74.90	1.94	0.213
DNB percentile rank in French	68.00	70.83	2.83	0.057
Predicted undergraduate major				
Major: STEM	0.392	0.378	−0.013	0.062
Major: selective STEM	0.170	0.181	0.011	0.347
Major: male-dominated STEM	0.277	0.274	−0.003	0.731
N	2,492	3,259	5,751	

Notes: Each row corresponds to a different linear regression with the dependent variable listed on the left for students enrolled in grade 12 (science track) in 2015/16. 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 an indicator that takes the value one if the school was visited by a professional and zero if the school was visited by a researcher, with the p-value reported in column 4. Standard errors are adjusted for clustering at the class level. Undergraduate majors are predicted for each student using the coefficients from a linear regression of the corresponding binary variable (e.g., enrolment in a STEM major) on all the school and student characteristics listed in the table. This model is fitted on the sample of students in the control group. [a] The *baccalauréat* pass rate is computed for students who were enrolled in grade 12 in 2014/15, i.e., in the year before the intervention, and who took the exams in the general or technical tracks. [b] The share of students enrolled in a STEM undergraduate major in higher education is computed for students who were enrolled in grade 12 (science track) in 2014/15.

Table E9 – Timing of Visits: Summary Statistics by Role Model Background

	All role models	Researchers (PhD/ Postdoc)	Professionals (employed by sponsoring firm)	Difference (3)−(2)	*p*-value of diff. (3)−(2)
	(1)	(2)	(3)	(4)	(5)
Panel A. Timing of classroom interventions					
November 2015	0.17	0.20	0.15	−0.05 (0.08)	0.51
December 2015	0.26	0.28	0.24	−0.05 (0.09)	0.60
January 2016	0.40	0.35	0.43	0.08 (0.10)	0.42
February 2016	0.17	0.15	0.19	0.04 (0.08)	0.62
March 2016	0.01	0.02	0.00	−0.02 (0.02)	0.32
Average nb of days since first visit (17 Nov 2015)	46.1	44.1	47.6	3.49 (6.04)	0.56
N	573	243	330		
Panel B. Time lag between intervention and student survey					
Average nb of days between visit and survey	67.6	71.8	64.5	−7.35 (6.31)	0.25
N	557	239	318		

Notes: Panel A reports the distribution of classroom visits by month of intervention and the average number of days since the first visit (17 November 2015). Panel B reports the average number of days between the classroom visit and the date of the student survey. The statistics are computed for all role models (column 1) and separately for researchers (column 2) and professionals (column 3). Column 4 reports the coefficient from the regression of each variable on an indicator that takes the value one if the classroom was visited by a professional and zero if the school was visited by a researcher, with the *p*-value reported in column 5. Standard errors (shown in parentheses) are adjusted for clustering at the role model × high school visit level. The date of the visit is missing for 7 out of the 98 participating schools, while the date of the survey is missing for 6 schools.

F Effects of Role Model Interventions: Additional Results

Table F1 – Perceptions of Science-Related Careers

	Girls			**Boys**			
	Control group mean	Treatment effect (LATE)	p-value [q-value]	Control group mean	Treatment effect (LATE)	p-value [q-value]	p-value of diff. (5)−(2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Grade 10							
Positive perceptions of science-related careers (index)	−0.020	0.245***	0.000	0.023	0.162***	0.000	0.013
		(0.027)	[0.001]		(0.027)	[0.001]	
Science-related jobs require more years of schooling	0.839	−0.087***	0.000	0.849	−0.075***	0.000	0.404
		(0.010)	[0.001]		(0.010)	[0.001]	
Science-related jobs are monotonous	0.290	−0.034***	0.003	0.318	−0.003	0.788	0.065
		(0.011)	[0.005]		(0.013)	[0.788]	
Science-related jobs are solitary	0.325	−0.057***	0.000	0.303	−0.059***	0.000	0.870
		(0.012)	[0.001]		(0.011)	[0.001]	
Science-related jobs pay higher wages	0.637	0.009	0.496	0.668	0.015	0.222	0.718
		(0.014)	[0.496]		(0.013)	[0.279]	
Hard to maintain work-life balance	0.297	−0.027**	0.014	0.283	−0.029***	0.009	0.916
		(0.011)	[0.018]		(0.011)	[0.016]	
N		6,475			5,751		
Panel B. Grade 12 (science track)							
Positive perceptions of science-related careers (index)	−0.003	0.296***	0.000	0.003	0.171***	0.000	0.002
		(0.032)	[0.001]		(0.033)	[0.001]	
Science-related jobs require more years of schooling	0.666	−0.113***	0.000	0.719	−0.096***	0.000	0.401
		(0.016)	[0.001]		(0.014)	[0.001]	
Science-related jobs are monotonous	0.169	−0.013	0.290	0.233	−0.022	0.185	0.615
		(0.012)	[0.291]		(0.017)	[0.185]	
Science-related jobs are solitary	0.228	−0.083***	0.000	0.206	−0.053***	0.000	0.080
		(0.012)	[0.001]		(0.013)	[0.001]	
Science-related jobs pay higher wages	0.531	0.063***	0.001	0.576	0.038**	0.018	0.327
		(0.018)	[0.002]		(0.016)	[0.030]	
Hard to maintain work-life balance	0.225	−0.046***	0.002	0.167	−0.015	0.174	0.092
		(0.015)	[0.003]		(0.011)	[0.185]	
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 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values associated with the synthetic index (highlighted in bold) are adjusted for multiple testing across the study's nine main outcomes of interest, separately by grade level and gender (see Appendix D for details). The q-values for the individual components of the index are adjusted for multiple testing across the index components, separately by grade level and gender. The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table F2 – Gender Differences in Aptitude for Mathematics

	Girls			**Boys**			
	Control group mean	Treatment effect (LATE)	p-value [q-value]	Control group mean	Treatment effect (LATE)	p-value [q-value]	p-value of diff. (5)−(2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Grade 10							
Equal gender aptitude for maths (index)	0.115	0.111*** (0.024)	0.000 [0.001]	−0.134	0.142*** (0.030)	0.000 [0.001]	0.383
M and W are born with different brains	0.211	−0.048*** (0.010)	0.000 [0.001]	0.209	−0.044*** (0.011)	0.000 [0.001]	0.742
Men are more gifted in maths than women	0.186	−0.028*** (0.010)	0.007 [0.007]	0.299	−0.048*** (0.014)	0.000 [0.001]	0.196
N		6,475			5,751		
Panel B. Grade 12 (science track)							
Equal gender aptitude for maths (index)	0.158	0.078*** (0.028)	0.004 [0.007]	−0.161	0.124*** (0.042)	0.003 [0.006]	0.348
M and W are born with different brains	0.143	−0.024** (0.011)	0.026 [0.026]	0.180	−0.032** (0.014)	0.027 [0.055]	0.618
Men are more gifted in maths than women	0.163	−0.028** (0.012)	0.021 [0.026]	0.266	−0.029* (0.016)	0.064 [0.064]	0.947
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 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values associated with the synthetic index (highlighted in bold) are adjusted for multiple testing across the study's nine main outcomes of interest, separately by grade level and gender (see Appendix D for details). The q-values for the individual components of the index are adjusted for multiple testing across the index components, separately by grade level and gender. The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table F3 – Taste for Science Subjects

	Girls			**Boys**			
	Control group mean	Treatment effect (LATE)	p-value [q-value]	Control group mean	Treatment effect (LATE)	p-value [q-value]	p-value of diff. (5)−(2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Grade 10							
Taste for science subjects (index)	−0.169	−0.033 (0.031)	0.275 [0.414]	0.197	−0.021 (0.026)	0.431 [0.555]	0.704
Enjoys maths (z-score)	−0.147	0.003 (0.030)	0.924 [0.924]	0.186	−0.005 (0.027)	0.844 [0.973]	0.818
Enjoys physics-chemistry (z-score)	−0.170	−0.042 (0.034)	0.222 [0.445]	0.223	−0.022 (0.029)	0.448 [0.896]	0.566
Enjoys earth and life sciences (z-score)	−0.042	−0.052 (0.037)	0.162 [0.445]	0.086	−0.027 (0.034)	0.438 [0.896]	0.492
Enjoys science in general	0.661	−0.011 (0.013)	0.384 [0.512]	0.790	−0.000 (0.011)	0.972 [0.973]	0.453
N		6,475			5,751		
Panel B. Grade 12 (science track)							
Taste for science subjects (index)	−0.002	0.018 (0.033)	0.583 [0.583]	0.002	0.014 (0.040)	0.733 [0.825]	0.924
Enjoys maths (z-score)	−0.097	0.086^{**} (0.036)	0.019 [0.076]	0.100	0.087^{**} (0.039)	0.027 [0.055]	0.976
Enjoys physics-chemistry (z-score)	−0.089	−0.005 (0.043)	0.911 [0.911]	0.102	−0.003 (0.040)	0.944 [0.945]	0.966
Enjoys earth and life sciences (z-score)	0.203	−0.040 (0.038)	0.288 [0.576]	−0.215	−0.070 (0.061)	0.246 [0.328]	0.603
Enjoys science in general	0.918	−0.002 (0.009)	0.770 [0.911]	0.930	0.022^{***} (0.008)	0.008 [0.034]	0.036
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 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values associated with the synthetic index (highlighted in bold) are adjusted for multiple testing across the study's nine main outcomes of interest, separately by grade level and gender (see Appendix D for details). The q-values for the individual components of the index are adjusted for multiple testing across the index components, separately by grade level and gender. The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table F4 – Self-Concept in Maths

	Girls			**Boys**			
	Control group mean	Treatment effect (LATE)	p-value [q-value]	Control group mean	Treatment effect (LATE)	p-value [q-value]	p-value of diff. (5)−(2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Grade 10							
Self-concept in maths (index)	−0.198	−0.001	0.981	0.231	0.033	0.250	0.324
		(0.028)	[0.982]		(0.029)	[0.375]	
Self-assessed performance in maths (z-score)	−0.127	−0.013	0.663	0.168	0.016	0.555	0.375
		(0.029)	[0.663]		(0.027)	[0.700]	
Lost in front of a maths problem	0.553	0.008	0.531	0.344	−0.005	0.700	0.437
		(0.013)	[0.663]		(0.012)	[0.700]	
Worried when thinking about maths	0.617	-0.029^{**}	0.025	0.420	-0.030^{**}	0.027	0.919
		(0.013)	[0.093]		(0.014)	[0.109]	
Can succeed in science subjects if puts in effort	0.843	0.018^{**}	0.046	0.883	−0.005	0.511	0.061
		(0.009)	[0.093]		(0.008)	[0.700]	
N		6,475			5,751		
Panel B. Grade 12 (science track)							
Self-concept in maths (index)	−0.184	0.051	0.139	0.187	0.068^{**}	0.038	0.695
		(0.035)	[0.157]		(0.033)	[0.057]	
Self-assessed performance in maths (z-score)	−0.126	0.044	0.202	0.123	0.080^{**}	0.017	0.387
		(0.034)	[0.270]		(0.034)	[0.035]	
Lost in front of a maths problem	0.486	-0.032^{*}	0.091	0.325	-0.032^{**}	0.050	0.980
		(0.019)	[0.183]		(0.016)	[0.067]	
Worried when thinking about maths	0.560	-0.037^{**}	0.044	0.384	-0.050^{***}	0.001	0.575
		(0.018)	[0.175]		(0.016)	[0.006]	
Can succeed in science subjects if puts in effort	0.942	−0.002	0.751	0.949	0.007	0.341	0.396
		(0.008)	[0.752]		(0.007)	[0.341]	
N		2,600			2,636		

Notes: This table reports estimates of the treatment effects of the role model interventions on students' self-concept in maths, 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 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values associated with the synthetic index (highlighted in bold) are adjusted for multiple testing across the study's nine main outcomes of interest, separately by grade level and gender (see Appendix D for details). The q-values for the individual components of the index are adjusted for multiple testing across the index components, separately by grade level and gender. The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table F5 – Science-Related Career Aspirations

	Girls			**Boys**			
	Control group mean	Treatment effect (LATE)	p-value [q-value]	Control group mean	Treatment effect (LATE)	p-value [q-value]	p-value of diff. (5)−(2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Grade 10							
Science-related career aspirations (index)	−0.103	0.005	0.851	0.120	0.004	0.871	0.977
		(0.029)	[0.970]		(0.027)	[0.872]	
Some jobs in science are interesting	0.845	0.018^{**}	0.042	0.854	−0.005	0.586	0.059
		(0.009)	[0.167]		(0.009)	[0.636]	
Would consider a job in science	0.466	−0.003	0.823	0.587	0.023^{*}	0.056	0.107
		(0.013)	[0.825]		(0.012)	[0.224]	
Interested in at least one STEM job[a]	0.642	0.003	0.825	0.849	0.009	0.332	0.671
		(0.012)	[0.825]		(0.009)	[0.636]	
Wage prospects important in career choice (z-score)	−0.045	−0.019	0.514	0.038	0.013	0.636	0.406
		(0.030)	[0.825]		(0.027)	[0.636]	
N		6,475			5,751		
Panel B. Grade 12 (science track)							
Science-related career aspirations (index)	−0.045	0.106^{***}	0.004	0.046	0.068^{*}	0.055	0.410
		(0.037)	[0.007]		(0.035)	[0.071]	
Some jobs in science are interesting	0.961	0.012^{**}	0.029	0.940	0.026^{***}	0.001	0.138
		(0.005)	[0.059]		(0.008)	[0.005]	
Would consider a job in science	0.721	0.023^{*}	0.078	0.762	0.038^{***}	0.006	0.404
		(0.013)	[0.104]		(0.014)	[0.012]	
Interested in at least one STEM job[a]	0.817	0.002	0.863	0.899	0.003	0.779	0.963
		(0.011)	[0.863]		(0.009)	[0.779]	
Wage prospects important in career choice (z-score)	−0.043	0.112^{***}	0.002	0.037	0.062^{*}	0.055	0.295
		(0.036)	[0.009]		(0.032)	[0.074]	
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 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values associated with the synthetic index (highlighted in bold) are adjusted for multiple testing across the study's nine main outcomes of interest, separately by grade level and gender (see Appendix D for details). The q-values for the individual components of the index are adjusted for multiple testing across the index components, separately by grade level and gender. The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. [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.

Table F6 – Grade 12 Students: Enrolment Status the Following Year (Detailed)

	Grade 12 (science track) students						
	Girls			Boys			
	Control group mean	Treatment effect (LATE)	p-value [q-value]	Control group mean	Treatment effect (LATE)	p-value [q-value]	p-value of diff. (5)−(2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. STEM undergraduate programs							
All undergraduate STEM majors							
Major: STEM	0.289	0.020	0.139	0.470	−0.002	0.925	0.310
		(0.014)	[0.157]		(0.019)	[0.926]	
Selective STEM majors							
Maths, physics, engineering, computer science (CPGE)	0.084	0.022^{**}	0.019	0.211	0.012	0.397	0.548
		(0.009)	[0.049]		(0.014)	[0.663]	
Earth and life sciences (CPGE)	0.020	0.007	0.172	0.010	0.001	0.768	0.293
		(0.005)	[0.272]		(0.003)	[0.769]	
Sciences - vocational (STS)	0.006	0.002	0.519	0.011	-0.005^{*}	0.099	0.092
		(0.003)	[0.520]		(0.003)	[0.247]	
Non-selective STEM majors							
Maths, physics, computer science	0.077	0.010	0.217	0.157	0.006	0.625	0.764
		(0.008)	[0.272]		(0.011)	[0.769]	
Earth and life sciences	0.103	-0.022^{**}	0.019	0.081	-0.016^{*}	0.064	0.620
		(0.009)	[0.049]		(0.009)	[0.247]	
Panel B. Non-STEM undergraduate programs							
All undergraduate non-STEM majors							
Major: non-STEM	0.507	-0.031^{**}	0.049	0.293	−0.008	0.571	0.286
		(0.016)			(0.015)		
Selective non-STEM majors							
Business and economics (CPGE)	0.021	0.001	0.826	0.017	0.006	0.220	0.453
		(0.004)			(0.005)		
Humanities (CPGE)	0.014	−0.002	0.584	0.003	−0.001	0.439	0.877
		(0.003)			(0.001)		
Other vocational (STS)	0.011	-0.009^{***}	0.002	0.008	-0.005^{**}	0.027	0.306
		(0.003)			(0.002)		
Non-selective non-STEM majors							
Medicine and pharmacy	0.259	−0.008	0.623	0.108	0.005	0.653	0.506
		(0.016)			(0.011)		
Law and economics	0.107	−0.006	0.580	0.079	−0.000	0.975	0.677
		(0.010)			(0.008)		
Humanities and psychology	0.080	−0.008	0.394	0.040	−0.007	0.265	0.924
		(0.009)			(0.006)		
Sports studies	0.018	−0.003	0.473	0.036	−0.005	0.441	0.814
		(0.004)			(0.006)		
Not enrolled in higher education	0.206	0.011	0.430	0.237	0.012	0.425	0.957
		(0.013)			(0.015)		
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' enrolment outcomes in the academic year following the classroom interventions, i.e., 2016/17, separately by gender. The enrolment 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 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values associated with the treatment effect estimates on the probability of enrolling in a STEM undergraduate major (panel A) are adjusted for multiple testing across the study's nine main outcomes of interest, separately by gender (see Appendix D for details). The q-values associated with the estimates for the different selective and non-selective STEM majors (panel A) are adjusted for multiple testing across these different STEM majors, separately by gender. The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table F7 – 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)	p-value of diff. (5)−(2) (7)
Obtained the baccalauréat	0.928	−0.014 (0.011)	0.176 [0.264]	0.877	−0.005 (0.010)	0.576 [0.577]	0.540
Baccalauréat percentile rank	53.54	−1.408* (0.780)	0.071 [0.214]	47.72	0.433 (0.760)	0.569 [0.577]	0.057
Baccalauréat percentile rank in maths	46.21	0.476 (0.845)	0.573 [0.574]	47.47	1.385 (0.901)	0.124 [0.373]	0.330
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 *baccalauréat* 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 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values are adjusted for multiple testing across the three *baccalauréat* outcomes, separately by gender. The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

G Robustness Checks

Table G1 – Treatment Effects on Student Perceptions: Estimates without Controlling for Baseline Characteristics

	Girls			Boys			
	Control group mean	Treatment effect (LATE)	p-value [q-value]	Control group mean	Treatment effect (LATE)	p-value [q-value]	p-value of diff. (5)−(2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Grade 10							
Positive perceptions of science-related careers (index)	−0.020	0.245*** (0.028)	0.000 [0.001]	0.023	0.167*** (0.029)	0.000 [0.001]	0.026
More men in science-related jobs	0.628	0.156*** (0.013)	0.000 [0.001]	0.629	0.168*** (0.014)	0.000 [0.001]	0.455
Equal gender aptitude for maths (index)	0.115	0.109*** (0.025)	0.000 [0.001]	−0.134	0.148*** (0.030)	0.000 [0.001]	0.273
Women do not really like science	0.157	0.059*** (0.011)	0.000 [0.001]	0.198	0.103*** (0.013)	0.000 [0.001]	0.003
W face discrimination in science-related jobs	0.603	0.127*** (0.013)	0.000 [0.001]	0.527	0.153*** (0.014)	0.000 [0.001]	0.123
Taste for science subjects (index)	−0.169	−0.038 (0.036)	0.294 [0.442]	0.197	−0.019 (0.031)	0.533 [0.685]	0.627
Self-concept in maths (index)	−0.198	−0.008 (0.031)	0.806 [0.807]	0.231	0.039 (0.032)	0.217 [0.326]	0.225
Science-related careers aspirations (index)	−0.103	0.012 (0.030)	0.695 [0.807]	0.120	0.007 (0.029)	0.801 [0.902]	0.906
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 [0.001]	0.003	0.155*** (0.033)	0.000 [0.001]	0.000
More men in science-related jobs	0.712	0.125*** (0.016)	0.000 [0.001]	0.717	0.149*** (0.015)	0.000 [0.001]	0.209
Equal gender aptitude for maths (index)	0.158	0.095*** (0.028)	0.001 [0.002]	−0.161	0.132*** (0.040)	0.001 [0.002]	0.447
Women do not really like science	0.074	0.044*** (0.009)	0.000 [0.001]	0.146	0.073*** (0.015)	0.000 [0.001]	0.089
W face discrimination in science-related jobs	0.624	0.095*** (0.020)	0.000 [0.001]	0.600	0.072*** (0.018)	0.000 [0.001]	0.344
Taste for science subjects (index)	−0.002	0.016 (0.034)	0.632 [0.633]	0.002	−0.000 (0.039)	0.998 [0.999]	0.721
Self-concept in maths (index)	−0.184	0.050 (0.039)	0.202 [0.228]	0.187	0.072** (0.035)	0.041 [0.062]	0.634
Science-related careers aspirations (index)	−0.045	0.113*** (0.037)	0.002 [0.003]	0.046	0.050 (0.033)	0.131 [0.169]	0.161
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 without 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 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 randomisation was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values are adjusted for multiple testing across the study's nine main outcomes of interest, separately by grade level and gender (see Appendix D for details). The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G2 – Treatment Effects on Enrolment Outcomes: Estimates without 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)	p-value of diff. (5)−(2) (7)
Panel A. Grade 10							
All STEM tracks							
Grade 11: science track	0.355	−0.004	0.753	0.551	−0.002	0.910	0.876
		(0.014)	[0.807]		(0.015)	[0.910]	
General versus technical STEM track							
Grade 11: science–general track	0.328	0.001	0.942	0.416	0.007	0.613	0.699
		(0.013)	[0.942]		(0.014)	[0.614]	
Grade 11: science–technical track	0.026	−0.005	0.128	0.135	−0.009	0.300	0.693
		(0.003)	[0.256]		(0.008)	[0.601]	
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	0.332
		(0.014)	[0.103]		(0.020)	[0.998]	
Selective versus non-selective STEM							
Major: selective STEM	0.110	0.035***	0.002	0.232	0.020	0.200	0.387
		(0.011)	[0.004]		(0.016)	[0.283]	
Major: non-selective STEM	0.178	−0.011	0.322	0.239	−0.017	0.212	0.745
		(0.011)	[0.322]		(0.014)	[0.283]	
Male- versus female-dominated STEM							
Major: male-dominated STEM	0.166	0.038***	0.002	0.379	0.017	0.387	0.287
(maths, 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	0.952
(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 the role model interventions on students' enrolment outcomes in the academic year following the classroom interventions, i.e., 2016/17, separately by grade level and gender, and without controlling for student baseline characteristics. The enrolment 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 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 randomisation was stratified by school. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (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 hypothesis testing, 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). The q-values associated with the treatment effect estimates on 'Grade 11: Science track' (panel A) and 'Major: STEM' (panel B) are adjusted for multiple testing across the study's nine main outcomes of interest, separately by grade level and gender (see Appendix D for details). The q-values associated with the treatment effect estimates for the different STEM tracks (panel A) or the different STEM majors (panel B) are adjusted for multiple testing across these different STEM tracks or majors, separately by grade level and gender. The p-value of the difference between the treatment effects by gender is reported in column 7. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

H Randomisation Inference

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

Randomisation 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. The intuition behind this approach is relatively straightforward. In RCTs, researchers know exactly how the randomisation was performed. Randomisation inference uses this knowledge to assess whether observed outcomes in a given sample are likely to have occurred 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.5]

In our setting, the ITT effect under the observed assignment to treatment is estimated using the following reduced-form specification:

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

where Y_{ics} is the observed outcome of student i in class c and school s, T_{cs} the observed treatment assignment of the student's class, $\boldsymbol{X}_{ics}$ the student characteristics in Table 1and θ_s the school fixed effects. The ITT estimate under the observed treatment assignment is denoted by $\hat{\beta}$.

To conduct randomisation inference, we proceed as follows. Taking into account the fact that randomisation 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.6] Let $\{P^r\}_{r=1}^R$ denote the set of R random placebo assignments from the randomisation 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:

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

where P_{cs}^r indicates assignment to a placebo treatment group for random draw r. School fixed effects, λ_s, are included to account for the fact that the randomisation 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 randomisation units (here classes), this method accounts for correlation within units. Following Davison and Hinkley (1997), we compute the p-values from a two-sided randomisation inference test of zero treatment effects as $p = (1 + \sum_{r=1}^R \mathbb{1}(|\hat{\beta}_r| \geq |\beta|))/(1 + R)$.

Table H1 presents the results of randomisation inference tests of the hypotheses that the role model interventions had no effect on student perceptions and enrolment outcomes, separately by grade level and gender. The ITT estimates $\hat{\beta}$ are shown in columns 1 and 4. The cluster-robust model-based p-values are reported and columns 2 and 5, while those based on randomisation inference are in columns 3 and 6. The results of the randomisation 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 enrolment in selective and male-dominated STEM programs for girls in grade 12.

[A.5] For more details on randomisation inference, see Rosenbaum (2010) and Imbens and Rubin (2015).

[A.6] See Paz and West (2019) for the number of draws to be used.

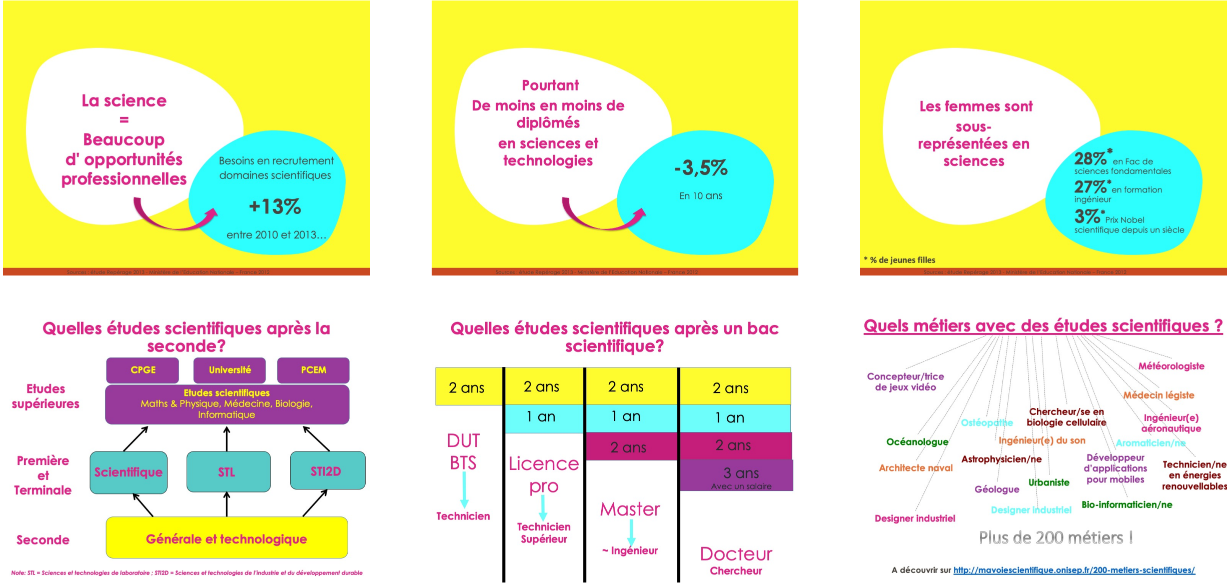
Table H1 – Randomisation Inference for Intention-to-Treat Estimates

	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)	0.227	0.000	0.000	0.151	0.000	0.000
More men in science-related jobs	0.143	0.000	0.000	0.158	0.000	0.000
Equal gender aptitude for maths (index)	0.103	0.000	0.000	0.132	0.000	0.000
Women do not really like science	0.052	0.000	0.000	0.094	0.000	0.000
Women face discrimination in science-related jobs	0.116	0.000	0.000	0.143	0.000	0.000
Taste for science subjects (index)	−0.031	0.280	0.320	−0.019	0.436	0.500
Self-concept in maths (index)	−0.001	0.981	0.980	0.031	0.255	0.320
Science-related career aspirations (index)	0.005	0.852	0.870	0.004	0.873	0.880
Enrolment outcomes						
Grade 11: science track	−0.002	0.863	0.880	−0.005	0.643	0.690
Grade 11: science–general track	0.003	0.796	0.810	0.004	0.713	0.740
Grade 11: science–technical track	−0.004	0.193	0.270	−0.009	0.240	0.300
N	7,241			6,459		
Panel B. Grade 12 (science track)						
Student perceptions						
Positive perceptions of science-related careers (index)	0.279	0.000	0.000	0.160	0.000	0.000
More men in science-related jobs	0.115	0.000	0.000	0.140	0.000	0.000
Equal gender aptitude for maths (index)	0.074	0.007	0.050	0.117	0.004	0.030
Women do not really like science	0.039	0.000	0.000	0.069	0.000	0.000
Women face discrimination in science-related jobs	0.080	0.000	0.000	0.070	0.000	0.000
Taste for science subjects (index)	0.017	0.592	0.710	0.013	0.738	0.810
Self-concept in maths (index)	0.048	0.150	0.300	0.064	0.042	0.140
Science-related career aspirations (index)	0.100	0.005	0.040	0.064	0.061	0.160
Enrolment outcomes						
Undergraduate major: STEM	0.019	0.154	0.310	−0.002	0.927	0.950
Undergraduate major: selective STEM	0.029	0.008	0.050	0.008	0.583	0.690
Undergraduate major: non-selective STEM	−0.010	0.341	0.510	−0.010	0.456	0.570
Undergraduate major: male-dominated STEM	0.031	0.005	0.030	0.012	0.495	0.610
Undergraduate major: female-dominated STEM	−0.014	0.175	0.350	−0.014	0.129	0.270
N	2,827			2,924		

Notes: This table presents the results of randomisation inference tests of the hypotheses that the program had no effect on student perceptions and enrolment outcomes. We randomly re-assign treatment 2,000 times among participating classes within each school and grade level, and re-estimate the ITT effects of these placebo treatments. The regression includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1 in the main text. 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 randomisation inference p-values are reported in columns 3 and 6. They are computed from a two-sided randomisation 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 $\mathbb{1}(\cdot)$ is the indicator function.

I Information, Persistence, Timing: Additional Results

I.1 Intensity of Information Provision



La science = Beaucoup d' opportunités professionnelles

Besoins en recrutement domaines scientifiques

+13%

entre 2010 et 2013…

Sources : étude Repérage 2013 - Ministère de l'Education Nationale – France 2012

Pourtant De moins en moins de diplômés en sciences et technologies

-3,5%

En 10 ans

Sources : étude Repérage 2013 - Ministère de l'Education Nationale – France 2012

Les femmes sont sous-représentées en sciences

28%* en Fac de sciences fondamentales

27%* en formation ingénieur

3%* Prix Nobel scientifique depuis un siècle

* % de jeunes filles

Sources : étude Repérage 2013 - Ministère de l'Education Nationale – France 2012

Quelles études scientifiques après la seconde?

Note: STL = Sciences et technologies de laboratoire ; STI2D = Sciences et technologies de l'industrie et du développement durable

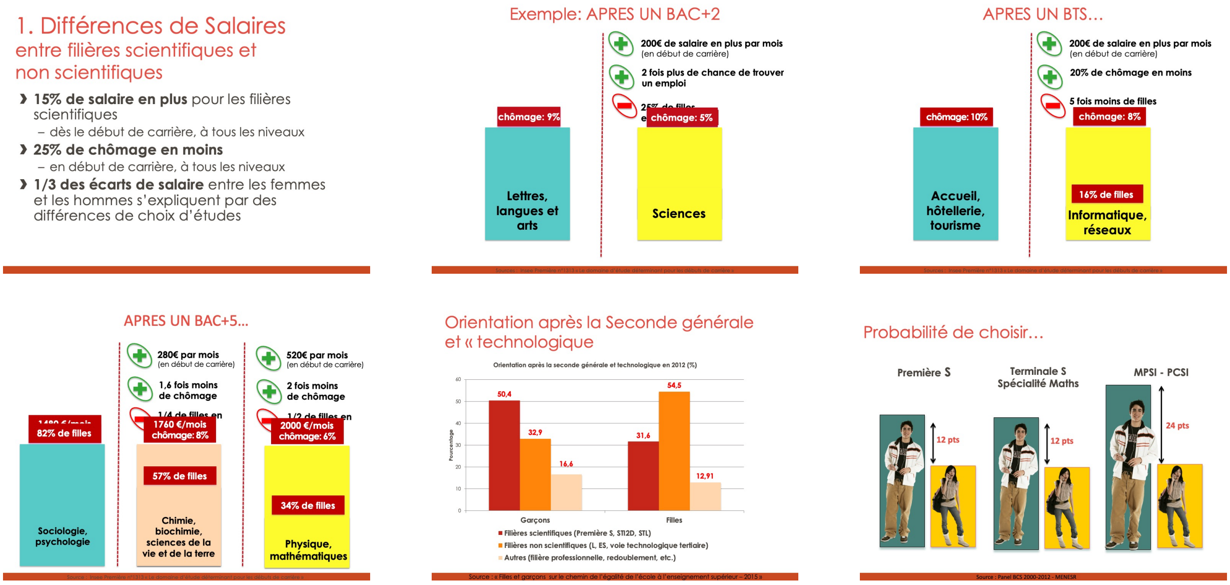
Quelles études scientifiques après un bac scientifique?

Quels métiers avec des études scientifiques ?

Plus de 200 métiers !

A découvrir sur http://mavoiescientifique.onisep.fr/200-metiers-scientifiques/

Figure I1 – Screenshots of the Slides Providing General Information on STEM Careers ('Regular Slides')



1. Différences de Salaires entre filières scientifiques et non scientifiques

- **15% de salaire en plus** pour les filières scientifiques
  - dès le début de carrière, à tous les niveaux
- **25% de chômage en moins**
  - en début de carrière, à tous les niveaux
- **1/3 des écarts de salaire** entre les femmes et les hommes s'expliquent par des différences de choix d'études

Exemple: APRES UN BAC+2

200€ de salaire en plus par mois (en début de carrière)

2 fois plus de chance de trouver un emploi

25% de filles e

Sources : Insee Première n°1313 « Le domaine d'étude déterminant pour les débuts de carrière »

APRES UN BTS…

200€ de salaire en plus par mois (en début de carrière)

20% de chômage en moins

5 fois moins de filles

Sources : Insee Première n°1313 « Le domaine d'étude déterminant pour les débuts de carrière »

APRES UN BAC+5…

280€ par mois (en début de carrière)

1,6 fois moins de chômage

1/4 de filles en

520€ par mois (en début de carrière)

2 fois moins de chômage

1/2 de filles en

Source : Insee Première n°1313 « Le domaine d'étude déterminant pour les débuts de carrière »

Orientation après la Seconde générale et « technologique

Orientation après la seconde générale et technologique en 2012 (%)

Pourcentage
60
50
40
30
20
10
0
50,4
32,9
16,6
31,6
54,5
12,91
Garçons
Filles
Filières scientifiques (Première S, STI2D, STL)
Filières non scientifiques (L, ES, voie technologique tertiaire)
Autres (filière professionnelle, redoublement, etc.)

Source : « Filles et garçons sur le chemin de l'égalité de l'école à l'enseignement supérieur – 2015 »

Probabilité de choisir…

Source : Panel BCS 2000-2012 - MENESR

Figure I2 – Screenshots of the Additional Slides Providing General Information on STEM Careers ('Augmented Slides')

Table I1 – Balancing Test: Classrooms Assigned to Role Models who were Provided with the Regular versus Augmented Sets of Slides

	Set of slides		Difference (2)−(1)	p-value of diff.
	Regular (1)	Augmented (2)	(3)	(4)
Panel A. Grade 10				
Student characteristics				
Female	0.532	0.526	−0.006	0.651
Age (years)	15.10	15.14	0.04	0.002
Non-French	0.052	0.067	0.015	0.014
High SES	0.396	0.369	−0.027	0.193
Medium-high SES	0.136	0.122	−0.014	0.059
Medium-low SES	0.238	0.244	0.006	0.559
Low SES	0.229	0.265	0.035	0.030
Number of siblings	1.467	1.500	0.033	0.383
Class size	33.86	32.76	−1.10	0.000
At least one science elective course	0.387	0.399	0.012	0.722
At least one standard elective course	0.746	0.759	0.012	0.648
DNB percentile rank in maths	58.37	58.57	0.20	0.886
DNB percentile rank in French	56.79	58.69	1.90	0.122
Predicted track in grade 11				
Grade 11: science track	0.443	0.457	0.014	0.248
Grade 11: science–general track	0.366	0.380	0.014	0.326
Grade 11: science–technical track	0.077	0.077	0.000	0.923
N	6,047	7,653	13,700	
Panel B. Grade 12 (science track)				
Student characteristics				
Female	0.491	0.492	0.001	0.951
Age (years)	17.12	17.13	0.01	0.673
Non-French	0.044	0.057	0.014	0.133
High SES	0.475	0.453	−0.022	0.519
Medium-high SES	0.140	0.132	−0.008	0.517
Medium-low SES	0.209	0.208	−0.001	0.936
Low SES	0.176	0.207	0.031	0.197
Number of siblings	1.479	1.516	0.037	0.454
Class size	32.13	31.83	−0.29	0.602
DNB percentile rank in maths	74.19	73.94	−0.25	0.870
DNB percentile rank in French	69.45	69.75	0.30	0.843
Predicted undergraduate major				
Major: STEM	0.384	0.382	−0.002	0.735
Major: selective STEM	0.179	0.175	−0.004	0.684
Major: male-dominated STEM	0.276	0.274	−0.001	0.856
N	2,748	3,003	5,751	

Notes: Each row corresponds to a different linear regression with the dependent variable listed on the left for students enrolled in grade 10 in 2015/16 (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 provided with the regular or augmented set of slides, respectively. Column 3 reports the coefficient from the regression of each variable on an indicator that takes the value one if the school was visited by a role model who received the augmented slides and zero if the school was visited by a role model who received the regular slides, with the p-value reported in column 4. Standard errors are adjusted for clustering at the class level. High school tracks in grade 11 are predicted for each student using the coefficients from a linear regression of the corresponding binary variable (e.g., enrolment in the general science track) on all the student characteristics listed in the table. This model is fitted on the sample of students in the control group.

Table I2 – Treatment Effects (ITT) for Grade 12 Students: Regular versus Augmented Slides

	Girls (1)	Boys (2)
Major: STEM		
Treatment group indicator (T)	0.024	−0.021
	(0.023)	(0.029)
T*Augmented slides	−0.006	0.029
	(0.037)	(0.040)
Major: selective STEM		
Treatment group indicator (T)	0.038***	0.021
	(0.014)	(0.024)
T*Augmented slides	−0.016	−0.018
	(0.021)	(0.035)
Major: male-dominated STEM		
Treatment group indicator (T)	0.048**	−0.003
	(0.019)	(0.030)
T*Augmented slides	−0.025	0.020
	(0.026)	(0.039)
Science-related jobs pay higher wages		
Treatment group indicator (T)	0.012	0.056***
	(0.032)	(0.021)
T*Augmented slides	0.087*	−0.055
	(0.049)	(0.036)
Positive perceptions of science-related careers (index)		
Treatment group indicator (T)	0.312***	0.173***
	(0.057)	(0.052)
T*Augmented slides	−0.042	−0.058
	(0.082)	(0.088)
Equal gender aptitude for maths (index)		
Treatment group indicator (T)	0.116***	0.046
	(0.041)	(0.061)
T*Augmented slides	−0.056	0.131
	(0.070)	(0.103)
N	2,827	2,924

Notes: This table reports estimates of the treatment effects (ITT) of the role model interventions on student outcomes for grade 12 students, separately by gender and by the type of slides (regular or augmented) that were provided to the female role model who visited the classroom. For each outcome of interest, the reported coefficients are obtained from a regression of the outcome on a treatment group indicator (T) and the interaction between this indicator and an indicator that takes the value one if the role model was provided with the augmented set of slides. The specification includes school fixed effects (to account for the fact that randomisation was stratified by school), month-of-visit fixed effects interacted with the treatment group indicator (to account for the fact that the additional slides were provided slightly later in the experiment) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (class). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

I.2 Persistence of the Effects and Timing of Visits

Table I3 – Persistence of Effects on Student Perceptions

	Girls			**Boys**		
	Days since intervention			Days since intervention		
	≤63 days	>63 days	p-value of diff.	≤63 days	>63 days	p-value of diff.
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Grade 10						
Positive perceptions of science-related careers (index)	0.289^{***} (0.037)	0.200^{***} (0.038)	0.083	0.162^{***} (0.037)	0.162^{***} (0.039)	0.999
More men in science-related jobs	0.146^{***} (0.016)	0.163^{***} (0.019)	0.479	0.188^{***} (0.019)	0.149^{***} (0.019)	0.138
Equal gender aptitude for maths (index)	0.143^{***} (0.031)	0.078^{**} (0.037)	0.182	0.173^{***} (0.041)	0.108^{***} (0.042)	0.258
Women do not really like science	0.080^{***} (0.015)	0.031^{**} (0.015)	0.022	0.114^{***} (0.016)	0.087^{***} (0.020)	0.292
W face discrimination in science-related jobs	0.139^{***} (0.017)	0.112^{***} (0.018)	0.262	0.165^{***} (0.018)	0.142^{***} (0.020)	0.379
Taste for science subjects (index)	−0.000 (0.045)	-0.068^{*} (0.040)	0.254	−0.028 (0.036)	−0.013 (0.037)	0.775
Self-concept in maths (index)	−0.050 (0.036)	0.051 (0.042)	0.063	−0.039 (0.042)	0.112^{***} (0.035)	0.005
Science-related career aspirations (index)	0.006 (0.038)	0.005 (0.042)	0.983	−0.023 (0.036)	0.035 (0.039)	0.274
N	3,119	3,356		2,856	2,895	
Panel B. Grade 12 (science track)						
Positive perceptions of science-related careers (index)	0.349^{***} (0.044)	0.249^{***} (0.044)	0.108	0.217^{***} (0.049)	0.125^{***} (0.043)	0.157
More men in science-related jobs	0.125^{***} (0.021)	0.120^{***} (0.023)	0.867	0.130^{***} (0.017)	0.167^{***} (0.026)	0.235
Equal gender aptitude for maths (index)	0.090^{**} (0.045)	0.068^{**} (0.032)	0.696	0.090 (0.059)	0.158^{***} (0.060)	0.420
Women do not really like science	0.072^{***} (0.015)	0.015 (0.011)	0.003	0.080^{***} (0.022)	0.067^{***} (0.021)	0.697
W face discrimination in science-related jobs	0.105^{***} (0.024)	0.068^{**} (0.030)	0.345	0.112^{***} (0.023)	0.038 (0.027)	0.038
Taste for science subjects (index)	−0.073 (0.049)	0.100^{**} (0.044)	0.010	0.048 (0.056)	−0.019 (0.057)	0.398
Self-concept in maths (index)	0.075 (0.049)	0.030 (0.049)	0.512	0.046 (0.037)	0.089^{*} (0.053)	0.504
Science-related career aspirations (index)	−0.021 (0.056)	0.221^{***} (0.041)	0.000	0.115^{***} (0.042)	0.022 (0.056)	0.182
N	1,201	1,399		1,255	1,381	

Notes: This table reports estimates of the treatment effects of the role model interventions on student perceptions, separately by grade level, gender and by the number of days between the date of the classroom visit and the date when students completed the survey. The sample is split at the median of this time interval, i.e., 63 days. On average, students below this threshold completed the survey 46 days after the intervention while those above completed it 93 days after, i.e., an additional 47 days. 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (class). The p-value of the difference between the treatment effects for students who took the before/after the 63 days threshold since the intervention is reported in columns 3 and 6, separately by gender. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table I4 – Effects on Enrolment Outcomes by Timing of Classroom Visits: Grade 12 Students

	Girls			**Boys**		
	Month of visit			Month of visit		
	Nov-Dec 2015 (1)	Jan-Feb 2016 (2)	p-value of diff. (3)	Nov-Dec 2015 (4)	Jan-Feb 2016 (5)	p-value of diff. (6)
Major: STEM	0.038*	0.003	0.210	0.054	-0.029	0.053
	(0.020)	(0.019)		(0.037)	(0.022)	
Major: selective STEM	0.046**	0.019	0.266	0.030	-0.002	0.324
	(0.019)	(0.015)		(0.028)	(0.017)	
Major: male-dominated STEM	0.038**	0.024	0.548	0.058	-0.011	0.098
	(0.018)	(0.016)		(0.035)	(0.021)	
N	1,253	1,461		1,257	1,575	

Notes: This table reports estimates of the treatment effects of the role model interventions on the enrolment outcomes of grade 12 students in the year following high school graduation, i.e., 2016/17, separately by gender and by whether the classroom visit took place before or after 31 December 2015. The enrolment outcomes are measured using student-level administrative data. 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (class). The p-value of the difference between the treatment effects for classroom visits that took place before versus after 31 December 2015, is reported in columns 3 and 6, separately by gender. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

J 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 participating schools. Section J.1 provides survey evidence suggesting that the scope for spillover effects was relatively limited. Section J.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.

J.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. 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 Table J1). In the treatment group, 58% of grade 10 students and 63% of science track grade 12 students report having talked about the classroom intervention with their classmates, but only 24% (27%) with schoolmates from other classes and 20% with students from other schools. Interestingly, these proportions are higher for girls than for boys: in grade 10, 66% of girls in the treatment group report having discussed the program with their classmates and 28% with schoolmates from other classes versus respectively 50% and 20% among boys; in grade 12, 70% of girls in the treatment group report having discussed the program with their classmates and 33% with schoolmates from other classes versus respectively 56% and 21% among boys.

In the control group, only 14% of students in grade 10 report having heard of classroom visits in other classes, mostly in a vague manner (12%). In grade 12, students in the control group are more likely to report being at least vaguely aware of such visits (34%), but less than 5% 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% versus 25%. 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.

J.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.7] Instead, our approach builds on

[A.7] Vazquez-Bare (forthcoming) develops a potential-outcome-based non-parametric framework to identify 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 J1, 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

$$Y_{icst} = D_s \cdot D_{cs} \cdot Y_{icst}(1, 1) + D_s \cdot (1 - D_{cs}) \cdot Y_{icst}(1, 0) + (1 - D_s) \cdot Y_{icst}(0, 0). \tag{A.3}$$

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.,

$$\overline{Y}_{s,t}(0, 0) = \frac{1}{N} \sum_{i=1}^{N} Y_{icst}(0, 0). \tag{A.4}$$

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

spillover effects in randomised 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 neighbours' observed treatment assignments.

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:

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

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

$$\Delta = \mathbb{E}\left(\frac{1}{M_1}\sum_{s\in\mathcal{S}_1}\left(\overline{Y}_{s,1}(1,0) - \overline{Y}_{s,1}(0,0)\right)\right). \tag{A.6}$$

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.,

$$\overline{Y}_{s,t} = \frac{1}{N}\sum_{i=1}^{N} Y_{icst}. \tag{A.7}$$

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.3), (A.4) and (A.7) imply that

$$\mathbb{E}(\overline{Y}_{s,t}) = \mathbb{E}\left(\overline{Y}_{s,t}(0,0)\right) \text{ if } s \in \mathcal{S}_0 \text{ and } t \in \{0,1\} \text{ or if } s \in \mathcal{S}_1 \text{ and } t = 0. \tag{A.8}$$

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, had all of its students 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 J1):

$$\widetilde{Y}_{s,1} = \frac{1}{N}\left(\sum_{i\in\mathcal{C}_s^0} Y_{ics1} + \left(1 + \frac{N_s^T}{N_s^C}\right)\sum_{i\in\mathcal{C}_s^C} Y_{ics1}\right), \quad s \in \mathcal{S}_1. \tag{A.9}$$

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.9), 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.3), the third equality follows from Assumption 1, while the last equality follows from Equation (A.5). 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 $\hat{\Delta}$:

$$\hat{\Delta} = \frac{1}{M_1}\sum_{s\in\mathcal{S}_1}(\widetilde{Y}_{s,1} - \overline{Y}_{s,0}) - \frac{1}{M_0}\sum_{s\in\mathcal{S}_0}(\overline{Y}_{s,1} - \overline{Y}_{s,0}). \tag{A.10}$$

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.10), 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.8) 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.6). □

Empirical specification. In the context of our study, the spillover effects estimator (A.10) 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. As a result, some of the schools that participated in the program evaluation in 2015, 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.9) 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:

$$\overline{Y}^*_{s,t} = \alpha + \theta_s + \theta_t + \sum_{k=2013}^{2015} \beta_k \cdot \mathbb{1}\{s \in \mathcal{S}_1 \text{ and } t = k\} + \epsilon_{s,t}, \qquad \text{(A.11)}$$

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 takes 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.8]

Selection of non-participating schools. To ensure that non-participating schools are as similar as possible to the participating schools, we use a nearest neighbour 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.9] 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.8] 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.

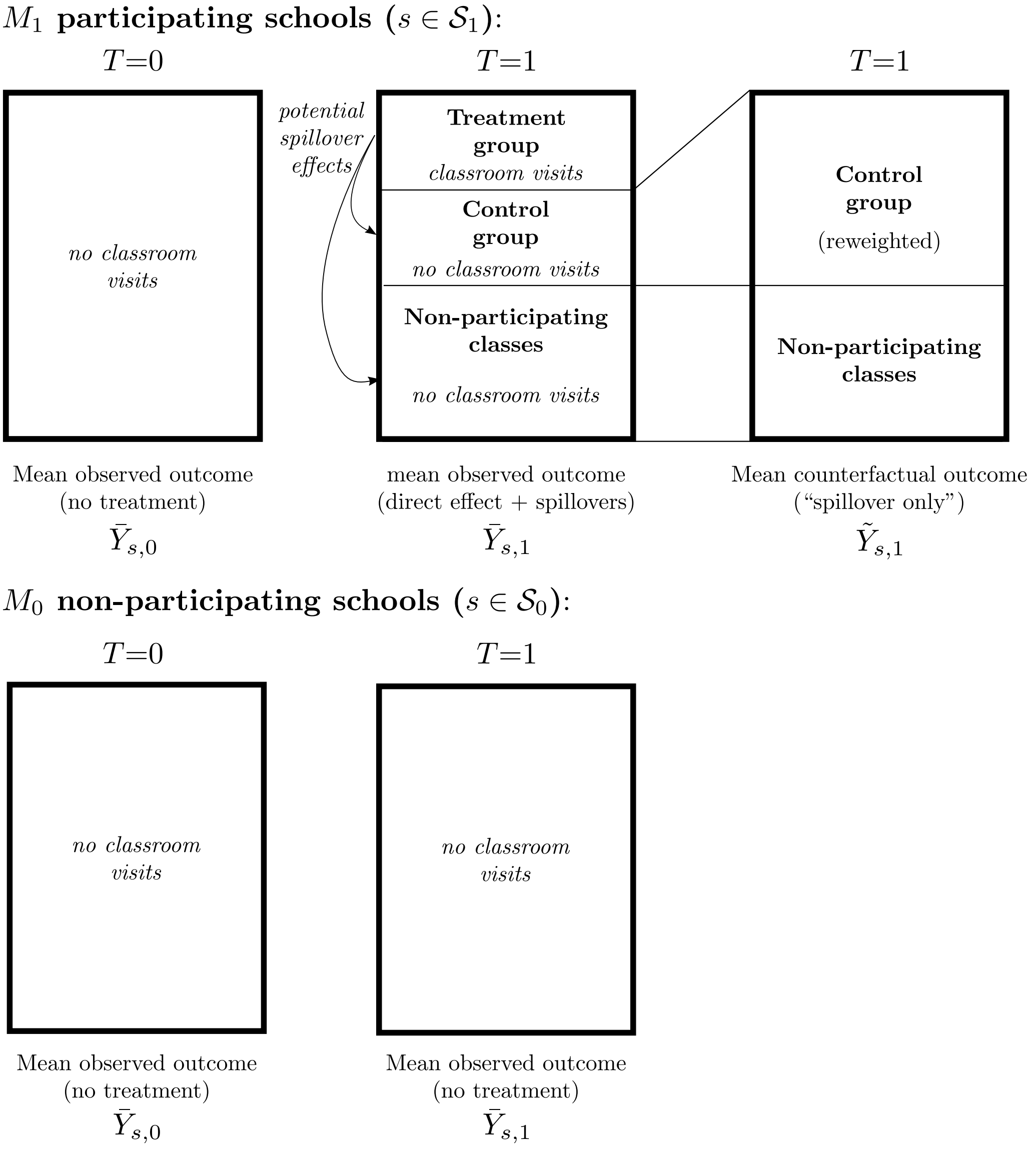
[A.9] The vector of exogenous school characteristics $\mathbf{X}_{st}$ includes the school's education district (Paris, Créteil or Versailles), whether it is public or private, and the following time-varying characteristics every year between 2012 and 2015: the number of students in grade 12 (science 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 enrolled in a STEM program after graduating from high school, the fraction who enrolled in a selective STEM program and the fraction who enrolled 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.11) to estimate the spillover effects of classroom visits on the college enrolment 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: enrolment in a STEM undergraduate program, enrolment in a selective STEM program and enrolment 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 J2. Panel A shows that the non-participating schools selected by the nearest-neighbour matching procedure are reasonably similar to the participating schools in terms of the average college enrolment 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 × t=2013) and on (participating school × 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 × $t = 2015$) 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 enrol in a selective STEM program, the comparison of treatment and control classes yields an estimated direct effect of $\hat{\beta} = 0.031$ (see Table 6 in the main text, column 2). Based on the results in column 2 of Table J2, the upper bound of the 95% confidence interval for the spillover effects is estimated to be $\hat{\Delta}^{UB} = 0.017$. Hence, in the case of selective STEM enrolment, we cannot reject spillover effects that would be at most 35% of the size of the 'true' direct effect $\hat{\beta} + \hat{\Delta}^{UB}$, which in this case would be of 4.8 percentage points. A similar calculation for the spillover effects on male-dominated STEM enrolment yields an upper bound of $\hat{\Delta}^{UB} = 0.025$. Since the estimated direct effect is $\hat{\beta} = 0.034$, we cannot reject spillover effects of at most 42% of the size of the 'true' direct effect $\hat{\beta} + \hat{\Delta}^{UB}$, which in that case would be of 5.9 percentage points.


M_1 participating schools ($s \in \mathcal{S}_1$):
T=0
T=1
T=1
no classroom visits
potential spillover effects
Treatment group
classroom visits
Control group
no classroom visits
Non-participating classes
no classroom visits
Control group
(reweighted)
Non-participating classes
Mean observed outcome (no treatment)
$\bar{Y}_{s,0}$
mean observed outcome (direct effect + spillovers)
$\bar{Y}_{s,1}$
Mean counterfactual outcome ("spillover only")
$\tilde{Y}_{s,1}$
M_0 non-participating schools ($s \in \mathcal{S}_0$):
T=0
T=1
no classroom visits
no classroom visits
Mean observed outcome (no treatment)
$\bar{Y}_{s,0}$
Mean observed outcome (no treatment)
$\bar{Y}_{s,1}$

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 J1 – 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 J1 – Scope for Spillover Effects: Summary Statistics from the Student Survey

	All	Boys	Girls	Within class Difference (3)−(2)	Within class *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 participating 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. The associated *p*-value is reported in column 5.

Table J2 – Difference-in-Differences Estimates of the Spillover Effects of Role Model Interventions on College Enrolment 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 versus 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)	(0.023)	(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-statistic	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., enrolment in a STEM undergraduate program, enrolment in a selective STEM program and enrolment 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 neighbour 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 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

K Heterogeneous Treatment Effects: Subgroup Analysis

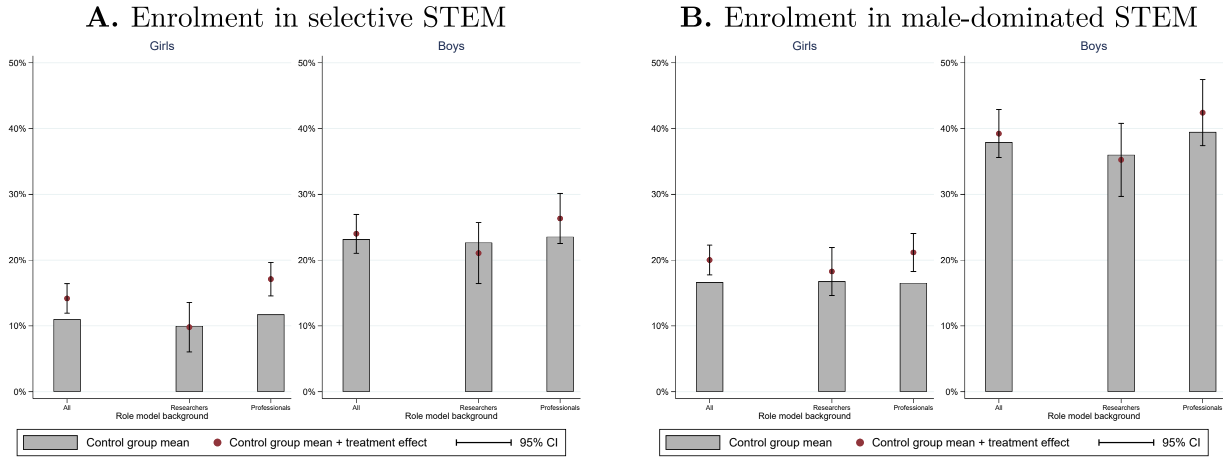


Figure K1 – Grade 12 Students: Enrolment in Selective and Male-Dominated STEM Undergraduate Programs, by Gender and Role Model Background

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 enrolment 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% 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors are adjusted for clustering at the unit of randomisation (class).

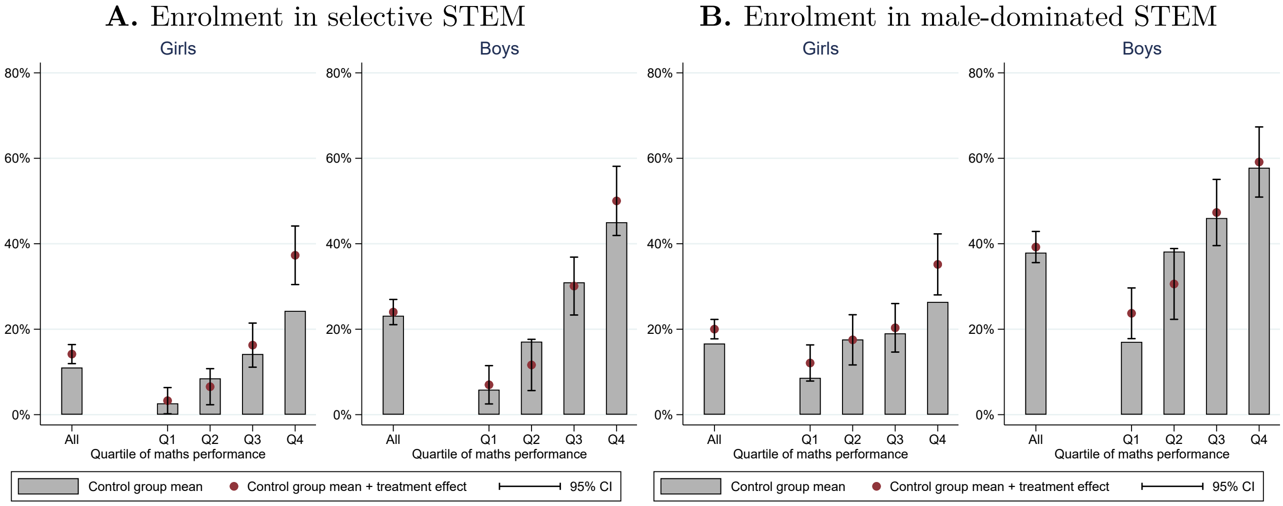


Figure K2 – Grade 12 Students: Enrolment in Selective and Male-Dominated STEM Undergraduate Programs, by Gender and Quartiles of *Baccalauréat* Performance in Maths

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 enrolment rates among students in the control group, both overall and separately by quartile of *baccalauréat* performance in maths. The solid circles show the estimated treatment effects (added to the control group means), with 95% 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 quartile of maths performance, using treatment assignment (interacted with the quartiles of maths performance) as an instrument for treatment receipt. The regression includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors are adjusted for clustering at the unit of randomisation (class).

Table K1 – Heterogeneous Treatment Effects on Grade 10 Students' Outcomes, 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)
Panel A. Enrolment Outcomes						
Grade 11: science track	0.005	−0.007	0.557	−0.028	0.012	0.102
	(0.015)	(0.015)	[0.956]	(0.018)	(0.017)	[0.307]
N	3,180	4,061		2,879	3,580	
Panel B. Student perceptions						
Positive perceptions of science-related careers (index)	0.225***	0.262***	0.474	0.119***	0.197***	0.152
	(0.038)	(0.036)	[0.956]	(0.043)	(0.033)	[0.342]
More men in science-related jobs	0.146***	0.160***	0.562	0.166***	0.172***	0.809
	(0.018)	(0.017)	[0.956]	(0.020)	(0.019)	[0.810]
Equal gender aptitude for maths (index)	0.056	0.155***	0.035	0.060	0.208***	0.015
	(0.035)	(0.033)	[0.317]	(0.048)	(0.037)	[0.080]
Women do not really like science	0.053***	0.059***	0.774	0.090***	0.111***	0.405
	(0.017)	(0.014)	[0.956]	(0.018)	(0.018)	[0.521]
W face discrimination in science-related jobs	0.123***	0.128***	0.851	0.136***	0.167***	0.244
	(0.020)	(0.016)	[0.956]	(0.021)	(0.017)	[0.367]
Taste for science subjects (index)	0.009	−0.067	0.213	−0.092**	0.036	0.018
	(0.045)	(0.042)	[0.956]	(0.039)	(0.036)	[0.080]
Self-concept in maths (index)	0.005	−0.005	0.864	0.010	0.051	0.473
	(0.045)	(0.037)	[0.956]	(0.041)	(0.039)	[0.533]
Science-related career aspirations (index)	0.004	0.007	0.956	−0.030	0.032	0.244
	(0.043)	(0.038)	[0.956]	(0.039)	(0.037)	[0.367]
N	2,933	3,542		2,608	3,143	

Notes: This table reports estimates of the treatment effects of the role model interventions on the outcomes of grade 10 students, separately by gender and by background of the female role model who visited the classroom (professional or researcher). 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 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 includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (class). Columns 3 and 6 report both the cluster-robust model-based p-value for the difference between the treatment effect estimates for students visited by a professional versus a researcher and, in square brackets, the p-value (q-value) adjusted for multiple hypothesis testing, 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 K2 – Heterogeneous Treatment Effects on Grade 10 Students' Outcomes, by Maths Performance

	Girls			**Boys**		
	Performance in maths			Performance in maths		
	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. Enrolment outcomes						
Grade 11: science track	−0.007 (0.014)	0.002 (0.017)	0.682 [0.897]	−0.017 (0.019)	0.006 (0.016)	0.364 [0.469]
N	3,584	3,484		3,221	3,075	
Panel B. Student perceptions						
Positive perceptions of science-related careers (index)	0.226*** (0.043)	0.262*** (0.036)	0.535 [0.897]	0.180*** (0.041)	0.146*** (0.039)	0.559 [0.630]
More men in science-related jobs	0.167*** (0.019)	0.142*** (0.017)	0.310 [0.699]	0.188*** (0.020)	0.152*** (0.017)	0.144 [0.325]
Equal gender aptitude for maths (index)	0.060 (0.037)	0.159*** (0.033)	0.047 [0.211]	0.105** (0.044)	0.178*** (0.042)	0.242 [0.363]
Women do not really like science	0.058*** (0.016)	0.054*** (0.014)	0.842 [0.897]	0.107*** (0.019)	0.096*** (0.017)	0.692 [0.692]
W face discrimination in science-related jobs	0.170*** (0.019)	0.084*** (0.017)	0.001 [0.008]	0.177*** (0.020)	0.131*** (0.019)	0.098 [0.325]
Taste for science subjects (index)	−0.036 (0.043)	−0.029 (0.038)	0.896 [0.897]	−0.071* (0.037)	0.029 (0.033)	0.032 [0.287]
Self-concept in maths (index)	−0.007 (0.038)	0.005 (0.037)	0.813 [0.897]	−0.006 (0.039)	0.070* (0.036)	0.122 [0.325]
Science-related career aspirations (index)	−0.032 (0.040)	0.040 (0.039)	0.186 [0.559]	−0.030 (0.041)	0.037 (0.035)	0.216 [0.363]
N	3,142	3,191		2,825	2,794	

Notes: This table reports estimates of the treatment effects of the role model interventions on grade 10 students' outcomes, separately by gender and performance in maths. Each row corresponds to a different linear regression performed separately by gender, with the dependent variable listed on the left. Students' performance in maths is measured from the grades obtained on the final maths exam of the *diplôme national du Brevet* at the end of middle school. Columns 1 and 2 (for girls) and columns 4 and 5 (for boys) report the LATE estimates for students below and above the median level of maths performance, 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 maths, using treatment assignment (interacted with the maths performance dummies) as an instrument for treatment receipt. The regression includes school fixed effects (to account for the fact that randomisation was stratified by school) and the student characteristics listed in Table 1. Standard errors (shown in parentheses) are adjusted for clustering at the unit of randomisation (class). Columns 3 and 6 report both the cluster-robust model-based p-value for the difference between the treatment effect estimates for students above versus below the median performance in maths and, in square brackets, the p-value (q-value) adjusted for multiple hypothesis testing, 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). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

L 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 L.1 gives an overview of the generic approach developed by Chernozhukov et al. (2018) to estimate, and make inference about, key features of heterogeneous effects in randomised experiments. Section L.2 provides further details on how we implement this method in the context of our study. Section L.3 explains how we extend this method to estimate the correlation between treatment effects. Finally, Section L.4 provides a detailed discussion of the results.

L.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 analyse 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, it can be implemented using any ML algorithm, allowing us to test algorithms of different degrees of sophistication, ranging from simple linear models to neural networks. Third, as described in Section L.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 characterise 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% 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% confidence interval for the final estimate as the median of the n 95% 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:

$$Y = \alpha_0 + \alpha B(Z) + \beta_1(T - p(Z)) + \beta_2(T - p(Z))(S(Z) - \mathbb{E}[S(Z)]) + \epsilon, \quad \mathbb{E}[w(Z)\epsilon X] = 0, \quad \text{(A.12)}$$

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.12) 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 if $s_0(Z) = s$. The main purpose of estimating β_2 is to check if the trained ML algorithms are able to detect heterogeneity.[A.10]

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 top and bottom 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 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.

[A.10] The intuition behind the formula for β_2 can be grasped by noting that Equation (A.12) 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 $s_0(Z)$.

L.2 Implementation of the Method

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

Population 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 enrolment 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 Table L2.

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.11] The sample splits are stratified by class, which is the randomisation 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, since 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 (Demirer, 2018).

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.12]

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

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

[A.11] 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.12] 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.

The above equation shows that maximising Λ is equivalent to maximising 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 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 maths tests of the exam, and a vector of 56 role model fixed effects.[A.13] 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 different treatment arms).

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

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

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

[A.13] 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%), we include indicators for missing grades as controls.

separate independent sample and are taken as given functions in Equation (A.13). The functions $p(Z)$ and $w(Z)$ and the vector X have the same meaning as in Equation (A.12). Equation (A.13) 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.13) identifies

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

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

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

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,[A.14] we can write

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

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

$$\rho_{A,B|Z} = \mathrm{Sign}(\beta_2^{A|B})\frac{\sqrt{\beta_2^{A|B}\beta_2^{B|A}}}{\sqrt{\beta_2^{B|B}}\sqrt{\beta_2^{A|A}}}. \tag{A.14}$$

Practical implementation. 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.14). 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

[A.14] 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 (A.14).

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.15]

The estimates based on Equation (A.14) 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 Table L7 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% confidence interval as follows.[A.16] At each iteration m of the data-splitting process, we compute $\hat{\rho}^{(m)}_{A,B|Z}$ (indexed by m) in the estimation sample. When $\hat{\rho}^{(m)}_{A,B|Z}$ can be computed, we estimate its 97.5% confidence interval using a clustered bootstrap procedure, which accounts for the 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% confidence interval of $\hat{\rho}^{(m)}_{A,B|Z}$ using the bootstrap percentile confidence interval method (Davison and Hinkley, 1997, chap. 5).[A.17] The 95% confidence interval for $\hat{\rho}_{A,B|Z}$ is then computed as the median of the 97.5% confidence intervals over the first 100 iterations for which $\hat{\rho}^{(m)}_{A,B|Z}$ could be computed—the price of the splitting uncertainty being reflected in the discounting of the confidence level from $1-\alpha$ to $1-2\alpha$.

L.4 Detailed Discussion of the Results

Heterogeneous treatment effects on enrolment outcomes. We use the procedure of Chernozhukov et al. (2018) described in Section L.1 to estimate the different features of the CATE on enrolment in selective or male-dominated STEM programs for girls in grade 12.

The machine learning results for girls in grade 12 are reported in Table L1. In panel A, the estimated ATEs of the interventions on grade 12 girls' enrolment in selective or male-dominated STEM are very close to those reported in Table 6 in the main text by virtue of the randomisation 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 enrolment 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 enrolment among grade 12 girls, GATEs ranging from a small negative effect in the bottom 20% to a large and significant 13.9 percentage point effect in the top 20%. The lesser heterogeneity in the effects on enrolment in male-dominated STEM is also confirmed, with no statistically significant difference between the top and bottom quintiles of treatment effects.

[A.15] For each pair of outcomes (Y^A, Y^B), Table L6 indicates the proportion of random data splits for which the correlation between CATEs could be computed.

[A.16] 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.

[A.17] The 97.5% confidence interval of $\hat{\rho}^{(m)}_{A,B|Z}$ is estimated using only the bootstrap samples for which $\hat{\rho}_{A,B|Z}$ can be computed.

Panel C describes the characteristics of the 20% most and least affected students (CLAN). The main takeaway is that the ML agnostic approach strongly confirms that the treatment effects on selective STEM enrolment are greater for high-achieving girls in maths and for those who were exposed to a professional rather than a researcher role model. Between the 20% most and least affected female students, the average gap in maths 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 enrolment 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% of girls with the largest treatment effects on selective STEM enrolment perform significantly better in French and are from higher socioeconomic backgrounds, compared with the least affected 20%. They are also less likely to have been exposed to role models who have children or who graduated in a male-dominated STEM field (maths, physics, engineering), and more likely to have been exposed to role models who participated in the FGiS program the year before. However, the fact that these characteristics are correlated both with students' maths 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.

Heterogeneous treatment effects on potential channels. The main results of the ML approach are reported in Table L3. 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 enrolment in selective STEM, i.e., student performance in maths and exposure to a role model with a professional background.[A.18]

The first key finding is that professionals and researchers were equally effective in debunking stereotypes on gender differences in maths aptitude, while they reinforced students' perceptions that 'women do not really like science' 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 emphasised men's and women's equal predisposition for science, cannot explain, alone, why the interventions increased girls' enrolment 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.

By contrast, Table L3 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 emphasising less the under-representation of women. Regarding perceptions of science-related careers, girls in the top quintile of treatment effects are 19.2 percentage points more likely to have been visited by a professional compared to girls in the bottom quintile, the difference being statistically significant at the 1% level. Professionals are similarly over-represented 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 under-representation in STEM: compared to the 20% of girls least affected for this outcome, the 20% most affected are 11.2 percentage points more likely to have been visited by a researcher.

The analysis of treatment effect heterogeneity by student maths performance tends to confirm that the messages conveyed by professionals were more effective in influencing female students' choice of study. Indeed, the students who were particularly receptive to these messages are also those for whom we find the strongest impact on STEM enrolment, i.e., high maths achievers.

[A.18] The heterogeneity loading parameter of the BLP and the GATEs associated with the best ML method are reported separately for each outcome in Table L4.

Average maths 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 under-representation in STEM and perception of gender discrimination increased the most.

While these comparisons on the basis of role model background and student maths 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.

Correlation between treatment effects. The correlations between treatment effects for girls in grade 12 are reported in Table L5, where the covariates that we use to predict treatment effect heterogeneity are the same as in Table L1. The results suggest that some channels were more important than others in steering female students towards STEM studies. The treatment effects on girls' enrolment 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 ($\hat{\rho} = 0.71$).[A.19]

While not statistically significant at the 5% level, the remaining correlations give some indication on the role of other candidate channels.[A.20] They confirm in particular that debiasing girls' attitudes towards gender differences in aptitude for maths is not associated with increased enrolment in selective STEM programs ($\hat{\rho} = 0.19$ with a 95% 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 enrol in such programs ($\hat{\rho} = 0.36$ $[-0.51, 2.01]$).

[A.19] The positive correlation between the treatment effects on taste for science and on enrolment in selective STEM suggests that students whose preferences were affected by the intervention also changed their choice of study. These effects, however, are highly heterogeneous (see Table L4): while the treatment effects on taste for science are positive for the 20% most affected girls in grade 12, they are negative for the 20% least affected, resulting in an average treatment effect close to zero (see Table F3).

[A.20] We report in Table L5 the lower and upper bounds for the lower and upper limits of the actual 95% confidence interval associated with each estimated correlation. Note that the (unknown) true confidence intervals are likely to be smaller than suggested by the bounds reported in this table.

Table L1 – Heterogeneous Treatment Effects on Selective and Male-Dominated STEM Enrolment 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)
Enrolment in selective STEM major				
Student characteristics				
Baccalauréat percentile rank in maths	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 PhD	0.692	0.606	−0.080	0.111
Field: maths, physics, engineering	0.316	0.226	−0.099	0.021
Field: earth and life sciences	0.618	0.602	−0.004	1.000
Enrolment in male-dominated major				
Student characteristics				
Baccalauréat percentile rank in maths	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 PhD	0.705	0.601	−0.103	0.043
Field: maths, 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 enrolment outcomes of 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 maths 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. 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. Panel C performs a Classification Analysis (CLAN) by comparing the average characteristics of the 20% 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.

Table L2 – Heterogeneous Treatment Effects on Selective and Male-Dominated STEM Enrolment 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)
Enrolment in selective STEM major				
Student characteristics				
Baccalauréat percentile rank in maths	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 PhD	0.707	0.664	−0.070	0.206
Field: maths, physics, engineering	0.359	0.236	−0.133	0.001
Field: earth and life sciences	0.541	0.688	0.157	0.000
Enrolment in male-dominated major				
Student characteristics				
Baccalauréat percentile rank in maths	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 PhD	0.660	0.682	0.020	1.000
Field: maths, 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 enrolment 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 maths 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. 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. Panel C performs a Classification Analysis (CLAN) by comparing the average characteristics of the 20% 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.

Table L3 – 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 maths	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 maths	74.87	25.00	−51.03	0.000
Class visited by professional	0.614	0.511	−0.112	0.031
Equal gender aptitude for maths (index)				
Mean baccalauréat percentile rank in maths	42.77	50.58	7.89	0.003
Class visited by professional	0.622	0.563	−0.058	0.403
Women do not really like science				
Mean baccalauréat percentile rank in maths	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 maths	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 maths	41.36	54.71	13.63	0.000
Class visited by professional	0.436	0.678	0.227	0.000
Self-concept in maths (index)				
Mean baccalauréat percentile rank in maths	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 maths	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 maths 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. The characteristics reported in this table are the students' average percentile rank in maths 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. The average treatment effects among the 20% most and least affected students can be found in panel B of Table L4.

Table L4 – 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 maths (index)	0.117	0.324	Random Forest
	[0.010]	[0.108]	
Women do not really like science	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]	
Self-concept in maths (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 maths (index)	0.019	0.246	0.210	Random Forest
	[1.000]	[0.037]	[0.332]	
Women do not really like science	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]	
Self-concept in maths (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 maths 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. 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.

Table L5 – Correlation between Conditional Average Treatment Effects (CATEs) for Girls in Grade 12

	Bivariate correlation with the CATE on enrolment in a selective STEM program	
	Estimate	95% confidence interval
	(1)	(2)
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 maths (index)	0.19	[−1.24, 2.05]
Women do not really like science	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]
Self-concept in maths (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 enrolment in a selective STEM program, denoted by $s_0^B(Z)$, and the CATE on each of the potential channels listed in the table, denoted by $s_0^A(Z)$. The proxy predictor of the CATE on selective STEM enrolment, 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. An indication of the quality of these predictions is provided by the heterogeneity loading (HET) parameter of the BLP (see Table L4, panel A). For each random split of the data, the correlation coefficient $\rho_{A,B|Z}$ is estimated as $\hat{\rho}_{A,B|Z} = \text{Sign}(\hat{\beta}_2^{A|B})(\hat{\beta}_2^{A|B}\hat{\beta}_2^{B|A})^{\frac{1}{2}}/(\hat{\beta}_2^{A|A})^{\frac{1}{2}}(\hat{\beta}_2^{B|B})^{\frac{1}{2}}$, 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 maths 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% 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.

Table L6 – 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 enrolment 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 maths (index)	0.35	1.00	0.98	0.36
Women do not really like science	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
Self-concept in maths (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 enrolment 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 $\hat{\rho}_{A,B|Z} = \text{Sign}(\hat{\beta}_2^{A|B})(\hat{\beta}_2^{A|B}\hat{\beta}_2^{B|A})^{\frac{1}{2}}/(\hat{\beta}_2^{A|A})^{\frac{1}{2}}(\hat{\beta}_2^{B|B})^{\frac{1}{2}}$, 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. Table L5 reports the median and 95% 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 L.3 of the Appendix.

Table L7 – Correlation between Conditional Average Treatment Effects (CATEs) for Girls in Grade 12: Sensitivity Analysis

	Bivariate correlation with the CATE on enrolment 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)	0.94	[0.20, 5.10]	0.73
More men in science-related jobs	−0.68	[−3.20, −0.01]	0.65
Equal gender aptitude for maths (index)	0.19	[−1.21, 1.97]	0.33
Women do not really like science	0.18	[−1.40, 2.70]	0.19
Women face discrimination in science-related jobs	−0.34	[−2.22, 0.56]	0.61
Taste for science subjects (index)	0.68	[0.07, 3.42]	0.66
Self-concept in maths (index)	−0.07	[−1.83, 1.40]	0.38
Science-related career aspirations (index)	0.36	[−0.52, 1.99]	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	[0.20, 4.89]	0.64
More men in science-related jobs	−0.66	[−3.15, −0.03]	0.62
Equal gender aptitude for maths (index)	0.16	[−1.18, 1.80]	0.28
Women do not really like science	0.18	[−0.87, 2.18]	0.05
Women face discrimination in science-related jobs	−0.31	[−2.17, 0.62]	0.59
Taste for science subjects (index)	0.59	[0.07, 2.48]	0.34
Self-concept in maths (index)	−0.07	[−1.71, 1.37]	0.28
Science-related career aspirations (index)	0.32	[−0.44, 1.78]	0.48

Notes: Similarly to Table L5, 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 enrolment in a selective STEM program, denoted by $s_0^B(Z)$, and the CATE on each of the potential channels 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% 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 L5.

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), 1481–1495.

Atelier Parisien d'Urbanisme (APUR), `DEPARTEMENT`*: Délimitation des 8 départements d'Île-de-France [database]*, Atelier Parisien d'Urbanisme, 2018. `https://opendata.apur.org/datasets/Apur::departement` (last accessed: 6 June 2020).

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.

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.

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.

Conférence des Grandes Écoles (CGE), *L'insertion des diplômés des Grandes écoles. Résultats de l'enquête 2018*, Conférence des Grandes Écoles, Paris, 2018. `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` (last accessed: 28 August 2019).

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

Demirer, Mert, `MLInference` *[R code]*, NBER Summer Institute 2018 presentation "Machinistas meet randomistas: useful ML tools for empirical researchers" by E. Duflo, 2018. `https://github.com/demirermert/MLInference/` (last accessed: 4 May 2018).

Direction des Études, de la Prospective et de la Performance (MENJ-DEPP), *Organisation des Concours et Examens Nationaux (OCEAN) [database]: OCEAN-DNB 2010–2015, OCEAN-BAC 2015 and 2016*, Ministère de l'Éducation Nationale et de la Jeunesse, 2017.

Délégation Académique à la Prospective et à l'Évaluation des Performances (DAPEP), *Base Élèves Académique (BEA) [database]: ELH 2012–2014, ELC 2013–2016, ELG 2015*, Rectorat de l'Académie de Versailles, 2017.

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.

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

Gayral-Taminh, Martine, Tomohiro Matsuda, Sylvie Bourdet-Loubère, Valérie Lauwers-Cances, Jean-Philippe Raynaud, and Hélène Grandjean, "Auto-évaluation de la qualité de vie d'enfants de 6 à 12 ans : construction et premières étapes de validation du KidIQol, outil générique présenté sur ordinateur," *Santé Publique*, 2005, *17* (2), 167–177.

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

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

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.

Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), *Enquête d'Insertion Professionnelle à 30 Mois des Diplômés de Master 2015 [database]*, Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, 2018. `https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-insertion_professionnelle-master_donnees_nationales/information/` (last accessed: 28 August 2019).

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.

Pôle Académique de la Prospective et de la Performance (PAPP), *Base Élèves Académique (BEA) [database]: ELH 2012–2014, ELC 2013–2016, ELG 2015*, Rectorat de l'Académie de Créteil, 2017.

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

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

Service Statistique de l'Académie de Paris (SSA), *Base Élèves Académique (BEA) [database]: ELH 2012–2014, ELC 2013–2016, ELG 2015*, Rectorat de l'Académie de Paris, 2017.

Sous-direction des Services d'Information et des Études Statistiques (MESRI-DGESIP/DGRI-SIES), *Système d'Information sur le Suivi de l'Étudiant (SISE) [database]: SISE-UNIV 2013–2016, SISE-ENS 2013–2016, SISE-INGE 2013–2016, SISE-MANA 2013–2016, SISE-PRIV 2013–2016*, Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI), Direction générale de l'enseignement supérieur et de l'insertion professionnelle (DGESIP), Direction générale de la recherche et de l'innovation (DGRI), 2017.

Vazquez-Bare, Gonzalo, "Identification and Estimation of Spillover Effects in Randomized Experiments," *Journal of Econometrics*, 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.