Gendered Teacher Feedback, Students' Math Performance and Enrollment Outcomes: A Text Mining Approach

Pauline Charousset, Marion Monnet

WORKING PAPER N° 2022-19

Gendered Teacher Feedback, Students' Math Performance and Enrollment Outcomes: A Text Mining Approach

Pauline Charousset
Marion Monnet

JEL Codes: I21, I24, J16
Keywords: teacher feedback, text mining, gender, student performance, higher education

Gendered Teacher Feedback, Students' Math Performance and Enrollment Outcomes: A Text Mining Approach*

Pauline Charousset

Marion Monnet

September 2025

Abstract

This paper examines how students' gender shapes teachers' feedback and its effects on performance and enrollment. Analyzing written feedback from math teachers to over 600,000 French high school students across five years, we find teachers praise girls for effort and positive behavior, while similarly-performing boys are more often criticized for behavior and praised for intellectual ability. Leveraging quasi-random teacher assignment, we estimate that gendered feedback slightly improves math performance, particularly for girls, especially when effort is emphasized. However, such feedback has no measurable impact on students' higher education enrollment in the following year.

JEL codes: I21, I24, J16.
Keywords: *teacher feedback*, *text mining*, *gender*, *student performance*, *higher education*

*Charousset: Paris School of Economics, 48 boulevard Jourdan, 75014, Paris, France (e-mail: pauline.charousset@ipp.eu); Monnet: Institute for Research in Education (Irédu) - Université Bourgogne Europe, 11 Esplanade Erasme, 21 000 Dijon, France (e-mail: marion.monnet@ube.fr). We are particularly grateful to Anne Boring, Alex Eble, Julien Grenet and Clémentine Van Effenterre for numerous reviews and wise advice. This paper also benefited greatly from discussions and helpful comments from Elliott Ash, Asma Benhenda, Philippe Colo, Etienne Dagorn, Marc Gurgand, Elise Huillery, Sylvie Lambert, Arnaud Maurel, Dominique Meurs, Roland Rathelot, Michael Stepner, Camille Terrier, and all the participants in the IPP Reading group, Sciences Po Education Policies 2020 seminar, Paris School of Economics Labor Chair Seminar 2021, ETH Zurich Data Science and Economics Seminar 2021, French National Institute for Demographic Studies Seminar 2021, Journées de Microéonomie Appliquée 2021, IWAEE 2021, University of Rennes Young Economist Seminar 2021, ASSA/AEA meetings 2022, PSE Young Economists of Education Workshop 2022, IZA Workshop in Economics of Education 2023, LEST seminar - Aix Marseille University, and the 10th Monash-Paris-Warwick-Zurich-CEPR Text-as-Data Workshop 2025. We are especially grateful to the French Ministry of Higher Education for access to their database. Financial support from the French National Research Agency (Agence Nationale de la Recherche) through project ANR-17-CE28-0001, EUR grant ANR-17-EUR-0001, Chaire Femmes et Sciences (Université PSL-Paris-Dauphine) and Chaire PEMS (Paris School of Economics) is gratefully acknowledged.

1 Introduction

Whether and how a student's gender influences the feedback he or she receives carries important implications for gender parity in education and in the labor market. Students' educational engagement and career choices depend significantly on the information about their ability that accumulates over their school career. This information is inherently noisy, in that it reflects not only intrinsic ability, but a whole range of other factors, such as assignment difficulty (Landaud et al., 2024), perseverance and effort (Alan et al., 2019), and others' perceptions of the student's performance (Sarsons, 2019). If teachers provide different kinds of feedback to male and female students, this could distort educational decisions, and possibly exacerbate gender differences in performance and career choices.

This paper provides an empirical test of whether student gender affects the feedback from math teachers in Grade 12 and examines how this relates to student performance and higher education enrollment choices. Using the entire set of Grade 12 student transcripts over the period 2013-2017, we analyze the written feedback from 6,716 math teachers to some 615,000 students in France and pay particular attention to the wording used by math teachers in assessing male or female students' performance. The features of our data further allow us to link the transcripts to national examination results, to assess how different feedback relates to student performance on these crucial, high-stakes tests. Our data also permit investigation of how gendered feedback correlates with other teacher characteristics or teaching practices, such as a measure of feedback personalization and grading bias. Last, we match our data with higher education application and enrollment data and follow students after high school graduation, comparing the educational careers of students exposed to different degrees of gendered feedback.

Exploiting this rich source of information, our first contribution is the analysis of written feedback in the light of gender differences. We propose a synthetic measure of gendered teacher feedback using text mining and machine learning techniques. We build a statistical model that predicts a student's gender based on the words used by his or her math teacher, controlling for gender differences in math performance. Comparing the prediction to students' actual gender, we then compute for each teacher the share of correct predictions, i.e. the accuracy of the model, which is our measure of gendered teacher vocabulary (GTV hereafter). The more a teacher uses female predictors to assess girls' performance and male predictors to assess boys', the better the predictive accuracy, and the more pronounced the teacher's gender differentiation.

Our first set of results provides evidence that math teachers do in fact provide differentiated feedback to equally good male and female students. The average GTV index is 64 percent

– that is, on average, our model correctly predicts the gender of 64 percent of students. By way of comparison, this is only marginally lower than the words' predictive power for students' performance, which can be taken as a sort of upper bound, given that the express purpose of the feedback is precisely to assess performance. However, we also document considerable variation in the distribution of GTV, suggesting that students are exposed to significantly variable degrees of gendered feedback.

To better understand how the vocabulary used by high-GTV teachers diverges from gender-neutral vocabulary, we perform a qualitative analysis of the words that best predict gender. Building on the psychology literature on teacher feedback and mindsets (Morgan, 2001; Burnett, 2002; Dweck, 2006), we classify them into five categories, reflecting different beliefs and expectations. First, words are classified as either positive, neutral, or negative. Second, they are classified as either "behavioral" or "competence"-related. The former are words referring to students' attitude in class or their effort; the latter relate to math concepts, the school environment or to students' intellectual ability.[1]

Our second set of results reveals marked gender differences in the kind of vocabulary used by math teachers to describe the work of equally able students. Two-thirds of the best female predictors are positive and mostly related to the student's behavior and effort, while two-thirds of the best male predictors refer to negative behavioral aspects. Positive male predictors, however, praise their intellectual skills. Overall, math teachers emphasize the positive behavioral aspects more heavily and encourage the effort of their female students, while equally performing male students are more severely criticized for their unruly behavior but tend to be praised for their intellectual skills.

We then relate our GTV measure to students' academic performance, higher education choices and enrollment outcomes. Using comprehensive national examination data, we seek to determine how exposure to a high-GTV math teacher affects students' grades on the national high school graduation exam (*baccalauréat*) in math. Higher education application and enrollment data further allow us to assess the effects of high GTV teachers on students' application behavior and their actual enrollment outcomes in the year following graduation. Our identification strategy exploits the variation in GTV between elective courses within a given high school, relying on the fact that teacher assignment to classes is practically as good as random conditional on a set of observable characteristics.

Our third result is that having a teacher with a 1-standard-deviation higher GTV is associated with an average gain in performance on the math *baccalauréat* exam of 1.4 percent of a standard

[1]Words that do not fit either of the two categories remain unclassified.

deviation; this effect is slightly (and significantly) greater for girls (1.8 percent vs. 0.95 percent of a standard deviation). The magnitude of the effect is admittedly moderate, but the impact on math performance on the *baccalauréat* is stronger for students who are exposed to teachers with an above-median GTV. Compared to students with teachers in the lowest decile of the GTV distribution, those with teachers in the fourth decile or above have math grades higher by up to 7.5 percent of a standard deviation. We also investigate whether teachers' vocabulary affects male and female students' math performance differently. We find that the positive effect of exposure to gendered feedback is reinforced when teachers are more likely to use the vocabulary associated with female students, i.e. emphasizing the positive behavior and effort. In keeping with the feedback and growth mindset literature, we find that females benefit more from the positive behavioral feedback than from the negative behavioral feedback and intellectual praise, while for males no difference between the effects of the two types of feedback is statistically detectable. The effects on students' top-ranked program in their college application and on their enrollment outcomes are very small and mostly not significant.

In the last part of the paper, we examine whether the effect of gendered feedback remains after controlling for a set of computable teaching characteristics. Since our empirical setting entails variation in exposure to teachers with differing levels of gendered feedback –rather than manipulation of the feedback content itself– we cannot rule out that GTV is correlated with other teacher attributes. We find that GTV is modestly correlated with gender grading bias and with a proxy for feedback personalization, measured by the textual distance distance across a teacher's written comments. Our main results however remain robust when controlling for both dimensions.

This paper contributes to several strands of the literature. It speaks first of all to the broad literature on performance feedback and individuals' beliefs and choices. This literature has focused on asymmetrical responses to performance feedback, seeking to determine whether individuals adjust their beliefs more after good or after bad evaluations (see for instance Zimmermann 2020; Coffman et al. 2024). In the educational setting, recent field experiments in economics have documented that performance feedback significantly affects academic investment, performance and enrollment decisions (Franco, 2019; Owen, 2022; Bobba and Frisancho, 2020), while other experiments (including psychological lab experiments) have focused on the nature of the feedback and the associated beliefs and expectations. In particular, this literature shows that assessments conveying the idea that intelligence is malleable (growth mindset) have a positive impact on students' motivation and attitudes (Corpus and Lepper, 2007), academic performance (Huillery et al., 2024), sense of belonging and willingness to pursue the subject (Good et al., 2012), while feedback implying that intelligence is innate (fixed mindset) has detrimental effects on those outcomes (Canning et al., 2021). These papers all investigate the impact of feedback in experimental settings (either laboratory or field experiment); ours aims at documenting the effect of feedback in a real-life setting.

Our paper also contributes to the social sciences literature that uses text as data to uncover patterns of gender biases and discrimination in various settings, including specialists' forums (Borhen et al., 2018; Wu, 2018), academia (Koffi, 2025), teaching material (Adukia et al., 2023), or the labour market (Ningrum et al., 2020). Our paper uses textual feedback as data to investigate whether the vocabulary employed by high school math teachers is gendered. We go beyond the description of gendered patterns, using the output from the textual analysis – GTV index – in a second step to relate it to students' educational outcomes.

Using non-textual data, another strand of the literature investigates how subjects' gender influences other people's perception of their ability and performance. Most of this work relies on proxy-matching techniques to investigate the differential treatment of observationally similar individuals, and the conclusion tends to be that women face higher evaluation standards than men in the labor market or academia (Sarsons, 2019; Sarsons et al., 2021; Card et al., 2021; Dupas et al., 2025). This paper, instead, uses a direct and comparable measure of ability to gauge how gender influences the assessment of individuals with similar objective performance levels.

Finally, our paper relates to the literature on the scope of gendered teacher behavior on student outcomes. Prior research provides evidence that teachers hold stereotyped beliefs about gender (Carlana, 2019), that they interact more with boys than with girls (Bassi et al., 2018), grade equally performing male and female students differently in their continuous assessment (Lavy and Sand, 2018; Terrier, 2020), and give them different career advice (Andersen, 2023). These gendered behaviors all have long-term consequences for academic schooling outcomes. We bring an original contribution to this strand of the literature by providing evidence of gendered behavior for another teaching practice, namely written feedback, and by documenting the short-run effect on student performance and higher education enrollment decisions of having teachers who follow different feedback practices.

The rest of the paper is organized as follows. Section 2 gives some institutional background on the French education system and the role of teacher feedback. Section 3 presents a conceptual framework and derives hypotheses on how exposure to gendered feedback affects performance and enrollment. Section 4 describes the various data sources and provides some descriptive statistics on the population of Grade 12 students and their math teachers. Section 5 presents our empirical strategy, detailing the steps taken in constructing our gauge of gendered teacher vocabulary (GTV), and measuring its impact on students' outcomes. Section 6 analyzes the gendered vocabulary in detail, with some statistics on the distribution of our GTV measure. Section 7 shows the impact of having a relatively high-GTV teacher on academic performance, higher education preferences and enrollment outcomes in the year following graduation. Section 8 examines how GTV correlates with other teacher attributes and assesses whether accounting for such characteristics affects the main results. Section 9 concludes.

2 Institutional Background

This section provides some background information on the French secondary education system as well as a brief presentation of the higher education application procedure. We also discuss the role of the written feedback in the French education system.

2.1 The French Secondary Education System

In France, secondary education consists of seven years of schooling: four years of middle school common to all students (*collège*, Grades 6 to 9), and three years of high school (*lycée*, Grades 10 to 12), which provide either vocational or general academic and technical training. Both the middle school and the high school curricula end with a national examination. At the end of middle school, students take the *Diplôme National du Brevet* (DNB), which tests their knowledge and skills in math, French and history and geography. At the end of Grade 11, high school students take the preliminary *baccalauréat* examinations, which include oral and written tests in French, as well as in history and geography for science major students. The remaining subjects are tested in the *baccalauréat* exam at the end of Grade 12. Only students who have earned the *baccalauréat* are eligible for higher education.

In general academic and technical high schools, after a common *Seconde générale et technologique* year (Grade 10), students are tracked into a general (80 percent of students) or a technical curriculum (20 percent of students). General academic track students further specialize by choosing a major at the start of Grade 11, and an elective course when they begin Grade 12. Students tend to specialize according to both their comparative advantage and their preferences, resulting in marked gender imbalances in majors and electives. Female students are slightly underrepresented among science majors (47 percent in 2018), while economics and humanities are largely female-dominated: 60 percent of economics majors and 80 percent of humanities majors in 2018 (MENJ-MESRI 2019).[2] These gendered patterns in choice of major are further reinforced by the choice of electives. The differences are particularly striking for science majors. Girls are strongly overrepresented in the earth and life science elective, where they account for 63 percent of students, against just 30 percent in computer sciences and 15 percent in engineering. The proportions are better balanced in the math and physics-chemistry electives (43 and 48 percent girls, respectively).

In any case, in French high schools, beyond the separation induced by the choice of an elective, gender segregation is limited. The composition of each class is determined by the principals, and, while students' electives are obviously taken into account, the principals also declare that gender is one of their top priorities (Cnesco, 2015). Most further say that they value some degree of heterogeneity in academic achievement levels, but in this area, unlike that of gender, stratification remains substantial. [3]

2.2 University Application and Enrollment

High school students apply to higher education programs in the Spring term of Grade 12. Throughout the year, the head teacher guides students with assistance in the application procedure and some counseling on choice of program. At the end of the academic year, the high school principal gives an opinion on students' chances of success in the programs listed in the application files, but students remain free to apply to whatever program they choose.

The higher education programs to which students can apply fall into two broad categories: non-selective university programs, which are open to all high school graduates, and selective programs. The latter include three different, academically stratified curricula: two-year undergraduate vocational and technical programs (*sections de techniciens supérieurs* and *instituts universitaires de technologie*), undergraduate management and engineering schools, and the two-year elite *classes préparatoires aux grandes écoles* (CPGE). The CPGE prepares students for the entry exam to the most prestigious French university programs (the *grandes écoles*) in science, business, or humanities.

Until 2017, the admission procedure for most undergraduate programs was centralized through the *Admissions Post-Bac* (APB) online platform.[4] The main step in the procedure

[2]We do not provide more recent figures given that our data stops in 2018 and that since then high school tracks have been reformed.

[3]Ly and Riegert (2015) inquired into the determinants of segregation within high schools, finding that the grouping of students by electives accounts for two-thirds of the observed social and academic segregation.

[4]In 2018, a major reform of the application procedure allowed universities to select students based on their past academic performance. Since our sample period is 2013-2017, the students considered here were not affected by this new system.

consisted in a variant of the college-proposing deferred acceptance mechanism (Gale and Shapley, 1962; Roth, 1982). Students were asked to submit a rank-order list of programs (ROL) that could include up to 36 choices, with a maximum of 12 choices per type of program (University program, STS, CPGE, etc.). After the submission deadline at the end of May, students were ranked by the different programs. For selective programs, the ranking was based on their students' academic records in Grade 11 and Grade 12: students' grades in the different subjects as well as teachers' written feedback were crucial role in the rankings of applicants. For non-selective programs, students were ranked according to a set of priority rules, based on their catchment area and the program's place in the student's list; that is, ranking was not based on students' grades.

2.3 Feedback in the French Education System

Each school year is divided into three trimesters, during which students are continuously assessed in each subject. At the end of each trimester, students receive a transcript detailing their GPA along with written feedback for each subject. These evaluations reflect teachers' perceptions of students' performance, work effort, and behavior in class, but also provide advice on how students can improve. This feedback serve as valuable and informative insights for students: it is discussed with teachers during class councils, shared with parents, and considered by selective higher education programs during the application process. Consequently, the way feedback is framed can significantly influence students' behavior and academic outcomes. In section 3, we outline the potential mechanisms through which these effects may occur.

3 Conceptual Framework: The Role of Feedback

The ultimate purpose of this paper is to investigate the impact of teacher feedback on two broad categories of student outcomes: their academic performance and their application and enrollment decisions. The effect of feedback on those dimensions is *a priori* unclear. Building on insights from the literature, we summarize potential mechanisms about how teacher feedback is expected to affect those two dimensions.

While the focal point of this conceptual framework mostly lies in examining the influence of teacher feedback, we acknowledge that, within our context, directly assessing the effects of teacher feedback on performance and enrollment choices is not feasible as we cannot manipulate the feedback that students receive. Instead, we utilize feedback more broadly as a proxy for gendered teacher behaviors (which may include discriminatory behaviours, gendered interactions with students, etc.).

3.1 Potential mechanisms regarding students' academic performance

Feedback may directly influence students' academic performance by providing information about what is required to succeed. Provided that the teacher's narrative regarding performance determinants is accurate, feedback may directly affect performance, and even more so as students usually operate under imperfect information.

Potential mechanism 1: Feedback may affect performance by providing information about what is needed to succeed.

Feedback is even more likely to positively impact performance when it carries information aimed at fostering a positive learning mindset. Such feedback is often categorized as effort-based or growth-mindset, emphasizing the process through which students can enhance their performance, in contrast to person-based or fixed-mindset feedback, which highlights students' inherent abilities (Dweck, 2006)[5]:

Potential mechanism 2: Teacher feedback is potentially more likely to enhance students' academic performance when it emphasizes effort over innate ability.

Finally, feedback may impact students' performance even if students do not fully endorse the teacher's narrative. In instances where feedback carries high stakes, students could strategically respond to their teachers' feedback even if they do not give credit to the information provided. This could be the case in our context, where the feedback we examine holds weight in higher education applications, as program officers consider it when evaluating applicants for selective programs. Consequently, students have an incentive to heed their teachers' guidance to increase their chances of receiving favorable feedback. This motivation is particularly pronounced in mathematics, which represents the most substantial workload for Grade 12 science track students and carries the highest weighting in their GPA and baccalauréat exam. The third potential mechanism can be summarized as follows:

[5] See for instance Blackwell et al. (2007); Delavande et al. (2019); Huillery et al. (2024) for the positive effects of specific growth-mindset interventions. Recent meta-analyses show that the average effect of growth mindset interventions on performance is relatively small (Macnamara and Burgoyne, 2023) but hide large heterogenous effects (Burnette et al., 2023).

Potential mechanism 3: Students may strategically respond to high-stakes feedback and improve their performance even when they do not give credit to the information provided.

3.2 Potential mechanisms regarding application and enrollment decisions

The information carried by teacher feedback may influence students' application and enrollment decisions by either altering their beliefs regarding their ability to succeed in a particular field, changing their identification with the subject, or both. There is general evidence suggesting that the nature and timing of feedback and information provision play a crucial role in determining the extent to which it can impact students' beliefs and choices. "Extraordinary" information and feedback, such as a one-time intervention during Grade 12, like role model class interventions or the disclosure of information on relative performance, can influence beliefs and application decisions (Breda et al., 2023; Hakimov et al., 2023). These studies indicate that even at a late stage of socialization, students' beliefs and aspirations are sufficiently malleable to respond to strategically provided information just before making application decisions.

Potential mechanism 4: Extra-ordinary information or feedback provided shortly before higher education applications may affect beliefs and applications decisions.

On the contrary, the information contained in the feedback we are considering can be viewed as "ordinary", as it represents just one of many feedback instances students have encountered throughout their academic history. However, it holds significance as it is the final feedback students receive before making their higher education applications, making it potentially pivotal. The impact of Grade 12 math teacher feedback thus hinges on how individuals process sequential and repeated feedback, as well as on how they weigh recent feedback against previous ordinary feedback. While in a Bayesian updating framework, individuals might not assign particular importance to the most recent feedback received, recent evidence suggests that the influence of feedback on beliefs and choices diminishes rapidly over time (Zimmermann, 2020; Coffman et al., 2024).

Potential mechanism 5: The last ordinary feedback is likely to exert the most significant effects compared to older ones but is unlikely to have long lasting effects on students' beliefs about their ability to succeed in or their identification with the field.

3.3 Potential mechanisms regarding the differential impact of feedback on male and female students

There is a consensus that male and female students react differently to feedback, with studies indicating that “uniform feedback can lead to significant performance losses and contribute to gender inequality” (Lovász et al., 2022). Thus, providing gendered feedback appears to be a rational approach for teachers. However, there is no agreement on the most effective type of feedback to enhance the performance of female and male students respectively. The efficacy of gender-tailored feedback depends crucially on contextual elements, particularly whether it is provided in a gender-stereotyped subject (Coffman et al., 2024).

Potential mechanism 6: Given that male and female students react differently to feedback, it could be rational for teachers to provide gendered feedback. However, the direction and the magnitude of the impact it may have on male and female students’ outcomes is a priori unclear.

4 Data and Summary Statistics

This section details the various data sources used to build our measure of gendered teacher vocabulary (GTV) and to measure the effect of exposure to a teacher with higher GTV (Section 4.1). We also present summary statistics on the sample of Grade 12 science majors and their math teachers (Section 4.2).

4.1 Data Sources

We use three main administrative databases: the higher education application data for six cohorts of Grade 12 students (2013-2017) collected via the APB platform, which includes detailed information on teacher feedback; the higher education enrollment data; and the data for the two main national exams (DNB and *baccalauréat*).

APB data. Our primary source is the comprehensive application data from the APB platform over the period 2013–2017. The platform collects a substantial amount of information in the application process. First, we use the students’ digitized academic records to retrieve teachers’ written feedback on all the subjects taken in Grade 12 (two trimesters). This is the main input for our measure of gendered teacher vocabulary. These transcripts also report the students' grades in the continuous assessment in both Grades 11 and 12. Teachers and students are uniquely identified, so the transcripts can be matched with the characteristics of students and teachers given in a separate APB file. Along with basic sociodemographic information (gender, place and date of birth, parents' socio-economic status, etc.), the APB data provide detailed information on high school careers (school track, major and elective choices), as well as information on the teachers' gender, subject taught, and head teacher status.

The APB data also record each applicant's final rank-order list of programs, the matching outcome (i.e. the program to which each student was admitted), and the students' acceptance decision (acceptance, conditional acceptance, rejection). We use this information to build our outcome variables.

School performance data. We use the OCEAN database, managed by the French Ministry of Education, to retrieve students' scores on two national examinations: the *Diplôme National du Brevet* (DNB), at the end of Grade 9, and the *baccalauréat*, at the end of Grade 12, both of which are graded anonymously and externally. The DNB serves as control for students' past academic performance in the estimation procedure; the *baccalauréat* is our main measure of student performance at the end of high school. To make the scores comparable across years, we transform the initial scores (ranging from 0 to 20) into percentile ranks, where 0 and 100 are the ranks for the lowest and the highest scoring students.

College enrollment data. To track Grade 12 students' subsequent enrollment outcomes, we use the *Système d'Information sur le Suivi de l'Étudiant* (SISE), managed by the Statistical Office of the French Ministry of Higher Education. This dataset, which covers the academic years 2013 to 2017, records all students enrolled in the French higher education system outside of CPGE and STS, except for the marginal number of students enrolled in undergraduate programs leading to paramedical and social care qualifications. For the selective programs, instead, we use a separate administrative data source, *Bases Post-Bac.*

Sample restrictions. Focusing on the feedback from math teachers, we restrict our sample to Grade 12 science majors, who naturally interact more frequently with their math teachers than humanities or social science majors.[6] These students are also the most likely to opt for a science major at university and may therefore be more responsive to their math teacher's

[6]The science major curriculum includes six hours of compulsory math classes (plus an extra two hours if the math elective is chosen) against four hours for the social science major (plus an hour and a half for the math elective) and none for the humanities major (four hours for the math elective).

feedback. We exclude students for whom the math teacher's identifier or the grade transcript is missing (45 percent of Grade 12 students in the science track in 2013, but diminishing to 11 percent in 2017; Table 1). In the vast majority of cases (between 68 and 93 percent of the missing observations), the teachers' identifiers and grade transcripts are missing because the high school as such was not reporting students' grades automatically on the APB platform. Dropping these observations is tantamount to dropping entire schools and accordingly does not threaten the internal validity of our analysis.[7] Finally, we restrict our sample to high schools with at least two science major classes, since our identification strategy relies on a within-school comparison of students (see Section 5), and to teachers who have taught at least two classes over the period 2013–2017. These restrictions remove between 11 and 25 percent of students. Depending on the year, the sample includes 49 to 72 percent of Grade 12 science major students, for a total of approximately 615,000 observations over the entire period.

[Insert Table 1 about here]

4.2 Summary Statistics

Students. Table 2 reports summary statistics of Grade 12 science majors' characteristics for the whole sample, separately for boys and girls. Students average 18 years of age and mostly come from high or a medium-high socio-economic background (43 and 16 percent respectively).[8] Girls are slightly underrepresented in the science major, making up 47 percent of science majors against 54 percent of all general academic track Grade 12 students (MENJS-MESRI, 2018). Turning to the elective courses, the gender differences are striking. Half of the girls opt for the earth and life science elective compared to only a quarter of the boys. Female students are also underrepresented in the math electives (19 percent vs. 27 percent) and engineering and computer sciences electives (6 percent vs. 20 percent). Another noticeable difference is in past academic performance, as measured by the national percentile rank on the DNB exam. Boys' average percentile rank in math is 3 points higher than girls'. On the other hand, females outperform males in French at both the DNB and *baccalauréat* exams by an average of 12 percentiles. Both

[7] It might, however, affect the external validity of our analysis. Table D5 in Appendix D shows the OLS coefficients of a dummy indicating whether all of a high school's grade transcripts are missing, regressed on the school's average characteristics. High schools with higher shares of high-SES, female and free lunch students are more likely to be reporting the grade transcripts. Reassuringly, the relative performance of female vs. male students at the math DNB examinations is related only marginally to the probability of not reporting grade transcripts.

[8] Students' socioeconomic status (SES) is measured by the Education Ministry's official classification, which uses the occupation of the child's legal guardian to define four groups: 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).

imbalances, in elective choices and past performance, are taken into account in our identification strategy and estimation procedure.

[Insert Table 2 about here]

Math teachers. Table 3 reports some descriptive statistics for the sample of math teachers in Grade 12 Science major courses. There are 6,716 teachers in the sample, of whom 57 percent are men. A little more than half served as head teacher of a class at least once during our sample period. Head teachers are likely to have a stronger influence on students' performance and enrollment behavior, in that they not only teach but also counsel their students. Each teacher is in charge of only one Grade 12 science major class on average each year, with an average class-size of 28 (90 percent of teachers teach only one Grade 12 science major class per year). Teachers appear nearly four times each in the 2013–2017 period, meaning that we have on average four classroom observations per teacher, which is crucial for the reliability of our GTV measure (see Section 5). Finally, the average length of teacher's feedback notes is 15 words, but with very considerable variability from teacher to teacher.

[Insert Table 3 about here]

5 Empirical Strategy

The first part of this section (Section 5.1) describes the estimation for measuring gendered teacher vocabulary (GTV). The second part (Section 5.2) presents the identification strategy to estimate the effect on students' outcomes of exposure to a teacher with higher GTV.

5.1 Measuring Gendered Teacher Vocabulary (GTV)

The measure of teacher GTV that we propose here leverages the rich data on teachers' written feedback to students in their Grade 12 academic records. To determine whether the words used to characterize a students' work, behavior and ability differ according to gender, we build a model that predicts the student's gender from the words that the teacher uses. Using machine learning techniques, we first estimate our model on a balanced subsample of Grade 12 science major students, controlling for class-level gender imbalances in students' previous academic performance. We then use this fitted model to compute a measure of gendered vocabulary for each teacher based only on the classes that he or she has taught. The different estimation steps, which draw on the text mining literature (Gentzkow et al., 2019), are presented below. The detailed procedure is described in Appendix A.

Data preparation. The first step was to convert the corpus of teacher feedback into a statistical database, itself done in two steps. First, using text mining techniques we replaced each word by its root so as to ensure its gender neutrality. Second, the corpus was converted into a matrix with one row per feedback and a number of columns equal to the number of distinct words appearing in the corpus (W_n). Each column is a dummy that takes value 1 if the word appears in the student's feedback, and 0 otherwise.

Student gender prediction. We assume that, conditional on the words used in the feedback, the probability of being a female student takes a logistic form:

$$P(Female_i = 1|W_i) = \frac{exp(\alpha W_i)}{1 + exp(\alpha W_i)} \quad \forall i, \tag{1}$$

Our objective is to find the set of α coefficients that minimize a penalized version of the negative log likelihood $\ln(L(\alpha))$ associated with Model (1), where λ denotes the regularization parameter chosen via a cross-validation procedure:

$$\hat{\alpha} = argmin_{\alpha}(-\ln(L(\alpha)) + \lambda \sum_{w=1}^{W_n} |\alpha_w|), \tag{2}$$

The model described by Equation (1) is trained on a subsample of Grade 12 students. Using the set of $\hat{\alpha}$ coefficients retrieved from the estimation procedure, we use the hold-out sample to predict each student's gender as follows:[9]

$$\widehat{P}(Female_i = 1|W_i) = \frac{exp(\hat{\alpha} W_i)}{1 + exp(\hat{\alpha} W_i)} \quad \forall i, \tag{3}$$

In practice, the model's predictive quality, or accuracy – the proportion of correctly classified observations – could be affected by two factors that we seek to neutralize before any estimation or prediction. First, since gender is correlated with math performance (see Section 4.1), Model (1) is likely to perform better on classes with stronger gender imbalances in math. To allay this concern, for each teacher we undersample, taking the same number of boys and girls from each quartile of prior math ability (proxied by DNB percentile rank). This ensures balanced training and hold-out samples, consisting of 50 percent male and 50 percent female students from each ability level. Second, feedback length varies substantially among teachers (see Table 3): obviously, a longer note, with more words used, is mechanically likely to generate more accurate predictions. For feedback of above-median length, we circumvent this issue by randomly sampling

[9]A student is classified as female if the predicted probability of being female is greater than 0.5 and male otherwise.

12 words (the median number).

Estimating Model (1) on the balanced subsample yields an accuracy of 63 percent: the model predicts the student's gender correctly in 63 percent of the cases. It performs slightly better at predicting male than female gender (65 vs. 63 percent).[10]

Gendered teacher vocabulary (GTV). We define each teacher j's GTV for class c as the share of students whose gender is correctly predicted. That is, GTV is the predictive accuracy of the model fitted on the teacher's sample of students. As teacher j's estimated GTV for class c might also capture some unobserved class-specific gender differences in behavior or performance, we compute an alternative measure that we call the *leave-one-out* GTV, defined as the average of teacher j's GTV over all the classes taught during the sample period, excluding class c. Our two measures are formally defined as follows:

$$GTV_{jc} = \frac{1}{N_{jc}} \sum_{i=1}^{N_{jc}} \mathbb{1}\{Sex_i = \widehat{Sex_i}\} \times 100 \quad \forall j, c, \tag{4}$$

where N_{jc} is the number of students in the balanced subsample of teacher j's students from class c, and:

$$GTV_{j \setminus c} = \frac{1}{N_j - 1} \sum_{c' \neq c} GTV_{jc'} \quad \forall j, c, \tag{5}$$

where N_j is the number of classes that teacher j taught throughout the period under study. In practice, both measures are computed as averages over 100 random balanced subsamples of teacher j's students.

Both measures range between 0 (the model systematically misclassifies females as males and males as females) and 100 (all students are assigned their actual gender). The greater the accuracy for a given teacher, the better we can recover their students' gender based on the words used in the feedback, and hence the stronger the gender differentiation in the vocabulary used in these assessments. A model that assigned student gender randomly with probability 0.5, would achieve a 50 percent accuracy. Thus, our model predicts gender better than random guessing for all teachers whose accuracy is above 50 percent[11].

[10]We tried more flexible specifications of the model by supplementing single words (unigrams) with interactions between words (bigrams), but this more complex specification did not improve predictive quality. We also tried more flexible and non-linear models such as random forests to predict gender based on the vocabulary dummies. Those models yield similar accuracies : 64%. We therefore kept the penalized logistic model as our preferred model given that its coefficients can be interpreted as odd-ratios.

[11]Less than 50 percent accuracy is possible in our setting, as the prediction is performed on small samples at the teacher level. However, averaging over a 100 estimations limits such random fluctuations, and the "leave-one-out" GTV is itself an average of multiple accuracies, reducing the noise inherent in the measure.

We compute two additional *leave-one-out* GTV measures to try and disentangle a teacher who is relatively more likely to use the "female vocabulary" from a teacher who is relatively more likely to use the "male vocabulary". For each teacher×class, we compute the share of correctly predicted female students on the one hand (the GTV-female measure), and that of correctly predicted male students on the other hand (the GTV-male measure). These two measures highlight different patterns in the vocabulary used by teachers. For example, a teacher for whom we predict his female students' gender very accurately and that of his male students poorly would have a high GTV-female and a low GTV-male. That is, such a teacher is very likely to use the female-specific vocabulary for girls and gender-neutral or even female-specific vocabulary for boys.

5.2 Identification Strategy

The second objective of this paper – in addition to documenting gendered practices in teachers' written feedback – is to characterize the relationship between our GTV measure and students' performance and enrollment outcomes. Our identification strategy compares students enrolled in the same high school, in the same elective course, but with math teachers having different levels of GTV. More specifically, we exploit the within high school× elective course × year variation in GTV and estimate the following equation:

$$Y_{isjcet} = \alpha + \beta_1 GTV_{j\backslash c} + \gamma_{set} + \epsilon_{isjcet}, \tag{6}$$

where Y_{isjet} is the outcome of student i in high school s with elective courses e taught by teacher j during academic year t. $GTV_{j\backslash c}$ is teacher j's standardized GTV measure and is class-specific, in that we use the *leave-one-out* GTV described in Equation (5). The coefficient γ_{set} is a set of school × elective course × year fixed effects. Hereafter GTV is standardized, and the coefficient of interest is β_1, which measures how a student's outcome is affected by a math teacher with a 1-standard-deviation higher GTV. The standard errors are robust and clustered at teacher level.[12] For this identification strategy to be valid, the *leave-one-out* GTV must not be systematically correlated with students' characteristics. We test this formally in Section 7.

[12]A preferable approach would be to bootstrap standard errors to account for prediction error, but owing to computational limitations we do not implement this correction.

6 Gendered Math Feedback

We first report the distribution of our GTV measures (Section 6.1) and then present a qualitative analysis of the gendered vocabulary used (Section 6.2).

6.1 The distribution of the GTV measures

Figure 1 shows the density of the overall *leave-one-out* GTV as well as the *leave-one-out* GTV measures separately for male and female students' predictions. There is evidence of a correlation between students' gender and the math feedback received, controlling for previously demonstrated math aptitude. Our model predicts gender better than random guessing for 90 percent of math teachers using the standard GTV measure and for over 95 percent using the *leave-one-out* GTV.[13]

For the median teacher, the student's gender is predicted correctly in 64 percent of the time. By comparison, the model achieves median accuracy of 66 percent in predicting the student's math performance, which is the upper bound of what we could expect given that the feedback serves precisely to assess this performance.[14] Breaking the GTV distributions down by teacher's gender, women math teachers differentiate their vocabulary slightly more, on average, than their male colleagues (see Figure C2).

Another observation that emerges from Figure 1 is that male students are more often correctly classified by our model as the *leave-one-out* GTV-male distribution is shifted to the right. On average males are correctly classified in 65 percent of cases against 63 percent for female students. Appendix Figure A1 further plots the correlation between GTV-male and GTV-female and shows that teachers for whom one gender is correctly predicted frequently have a substantially lower proportion of correct classifications for the other gender. A 1-standard-deviation increase in GTV-male is associated with a 0.6-standard-deviation decrease in GTV-female. This suggests that teachers using the female-specific vocabulary for their female students are not systematically using the male-specific vocabulary for males, but rather a female or gender-neutral vocabulary.

[Insert Figure 1 about here]

[13]Overall, only 4 percent of the teachers have a leave-one-out GTV below 50 percent, nearly all of these scoring between 44 percent and 50 percent. This is explained largely by the fact that these teachers are observed only three times on average, compared to 4 times for other teachers. Their leave-one-out GTV is therefore somewhat noisier.

[14]For performance prediction, we reestimate equation 1 with the response variable being equal to 1 if the student is among the top 50 percent performers of their class at the math DNB exam and 0 otherwise.

6.2 Qualitative Analysis of the Best Gender Predictors

Definition of the classification. A high degree of gender differentiation in the vocabulary used (i.e. a high GTV), may express differing beliefs and expectations on the part of teachers. It may reflect gender stereotypes on students' math aptitude, but it could also be that the teacher adapts feedback to the different student profiles if they perceived that male and females students react differently to feedback, as highlighted in the conceptual framework section 3.3.

To gauge the extent to which the gendered vocabulary expresses gendered beliefs and expectations, we explore the actual feedback content by analyzing the best gender predictors. Building on the psychology literature on the classification of teacher feedback and mindsets (Morgan, 2001; Burnett, 2002; Dweck, 2006), we classify the gender predictors into five different categories to capture the different beliefs and expectations conveyed. First, depending on the valence of the word, we categorize it as positive, neutral or negative. Second, words referring to students' attitude in class or the effort dedicated to the subject are classified as "behavioural", while those relating to math concepts, the school environment or to the students' intellectual ability are classified as "competence-related". Words that do not fit either category remain unclassified.[15] The classification of the top 100 male and female predictors is shown in Appendix Tables B2 and B3.[16]

Analysis of the best gender predictors. The analysis reveals marked differences in the qualifiers used by teachers according to the student's gender. Figure 2 reports the odds ratios derived from the estimation of the model described by Equation (1) for the top 10 predictors of each gender. Feedback referring to lack of confidence, propensity to get discouraged or cheerful aspect ("smiling") is between 1.7 and 2.1 times more likely to be directed to a girl than to a boy, relative to other feedback. Teachers are also more likely to note that female students are panicked, and to cite their exemplary conduct ("exemplary", "studious"). On the other hand, feedback that describes the student as childish ("has fun"), comments on the lack of organization ("messy"), or praises the student's curiosity, passion and intuitions is between 1.8 and 2.6 times more likely to be addressed to a male than to a female student, compared to

[15] We sought to confirm our classification with data-driven techniques using bi-term topic models tailored for short texts. However, these models performed poorly on our data, which are highly specific, in that the texts are very short, averaging just 15 tokens per teacher, and the overall vocabulary is quite limited (on average 1,600 words), with little variation in the topics (almost all relating to academic performance and behaviour). This presented the typical challenges inherent in such short texts: the topics generated gathered inconsistent words (*trivial topics*) and the different topics were highly similar with a large share of words in common (*repetitive topics*, see Wu et al. 2020 for a discussion of these issues.)

[16] Every token has been classified, but we show only the top 100 predictors, insofar as the others are not used more frequently for girls or boys (their odds ratio is around 1) and in the vast majority of cases cannot be classified in any of the five categories.

feedback that does not mention these terms.

[Insert Figure 2 about here]

Figure 3 extends the analysis to the 30 best predictors of each gender and plots them on a quadrant that distinguishes positive from negative words (neutral words being in the middle), where the marker symbols refer to competence, behavioural or unclassified words. The first striking feature is the relative proportions of positive versus negative feedback by gender. Among the top 30 male predictors, only about one-third are positive against two thirds for the best female predictors. Most interestingly, conditional on being positive, almost all the best male predictors are competence-related ("curious", " idea", "interest", "intuition", "passion"), while nearly all the best female predictors refer to behavioural aspects ("diligent", "willingness", "studious"). On the other hand, two-thirds of the best male predictors can be classified as negative, referring overwhelmingly to disruptive behaviour ("has fun", "childish") or to lack of work-effort ("waste", "superficial"). Negative female predictors (one-third) underline their lack of confidence and their stress but also, to a lower extent, their difficulties in math.

[Insert Figure 3 about here]

Analyzing the top 100 gender predictors together, the foregoing results stand confirmed. Figure 4 Panel A shows the proportions of negative, positive and neutral feedback, conditional on whether it is competence-related or behaviour-related. Panel A shows that only 7 percent of the female predictors that can be classified as competence-related are positive, compared with 50 percent for male predictors, while 33 percent are negative (7 percent for male students), and the rest are neutral. Symmetrically, among the female predictors classified as behavioural, 63 percent correspond to positive feedback compared with 20 percent for males. The latter receive a much larger share of negative feedback: as much as 71 percent of behavioural male predictors are negative, as against just only 22 percent for females.

Turning next to the proportions of competence, behavioural and neutral words used in positive and negative feedback (Panel B), we find that conditional on being positive, the top female predictors are competence-related in only 3 percent of cases compared with 54 percent for their male counterparts. For negative predictors, 64 percent of the male predictors and 46 percent of the female predictors are related to behavioural matters, for negative competence-related predictors these proportions are respectively 6 percent and 25 percent.

[Insert Figure 4 about here]

All in all, these descriptive statistics indicate that teachers do use a differentiated vocabulary for their male and female students. They seem to insist more on positive behavioural aspects and to encourage effort with their girl students, while equal performing males are more likely both to be criticized for unruly behaviour and to be praised for intellectual skills. In the following section, we investigate how this gendered feedback affects students' performance, future choices and enrollment outcomes.

7 The Impact of Gendered Feedback on Student Outcomes

Having found significant differences in the gendered vocabulary of Grade 12 math teachers, we now turn to their effect on students' outcomes. First, we report a series of statistical tests to validate our empirical strategy (Section 7.1). We then discuss how the exposure to teachers with different levels of GTV affects academic performance, higher education application behavior and enrollment in the year after high school graduation (Section 7.2). Our results prove to be robust to a series of alternative specifications.

7.1 The Validity of the Empirical Strategy

7.1.1 Exogeneity Assumption

For our identification strategy to be valid, teacher GTV cannot be systematically correlated with students' characteristics. Ideally, we would want teachers to be assigned randomly to classes within a school for a given elective course. We test this formally below.

Balancing tests. The coefficients from a regression of teachers' standardized *leave-one-out* GTV, defined at the class level, on students' socio-economic characteristics and baseline academic performance agregated at the class level, along with a set of school$\times$elective course$\times$year fixed effects, are reported in Table 4. Teacher GTV is not systematically correlated with students' observable characteristics. Of the twelve characteristics included in the regression, none are statistically significant. The test tells in favor of the random allocation of teachers conditional on school$\times$elective course$\times$year fixed-effects.

[Insert Table 4 about here]

Random allocation of students. To check whether students are allocated randomly to teachers within a given high school, elective course and year, we perform a series of Pearson's Chi-square tests of independence. For each unique combination of school, elective course, and year, we tabulate math teachers' identifiers with each of the students' baseline characteristics and test for independence.[17] Table 5 reports the percentage of p-values below the nominal values of 0.05 and 0.01. Except for the female dummy and the age variable, we find that the empirical p-values are close to the nominal values (between 0.4 percent and 4.3 percent of p-values are below the nominal levels, i.e. signifant at the 5% level). For the female dummy, the empirical p-value is 7.6 percent (respectively 5.6 percent for the age variable). That is, in about 8 percent of every 100 high school×elective×year combinations, we cannot exclude the non-random assignment of female students to classes at the 95 percent level. To ensure that the results in Section 7.2 are not driven by these slight gender imbalances, Equation (6) is also estimated with the average proportion of females in the class as an additional control, as well as with the full set of students' baseline characteristics.

[Insert Table 5 about here]

Overall, the tests indicate that in any given school, elective course and year, students' allocation to classes is close to random.

7.1.2 Reverse Causality

A legitimate concern regarding the GTV measure is that teachers' behavior could be influenced by the type of students in their class. In this case, the measure would not be picking up any stable trait in the teachers' gendered vocabulary. However, we show that this type of reverse causality is unlikely to be an issue in our setting.

To begin with, students' observable characteristics are rather well balanced across the distribution of teacher GTV (see Table 4): teachers who differentiate their vocabulary more or less are not systematically assigned any particular type of students.

Second, to each class we assign its teacher's *leave-one-out* GTV measure, i.e., the average GTV measured in all the other classes ever taught by that teacher, so that students have no effect on the GTV measure they are being assigned.

Third, to determine whether gendered vocabulary is specific to math teachers or to classrooms' characteristics, we replicate the GTV estimation procedure for other Grade 12 science track

[17] Continuous baseline characteristics such as age are previously dichotomized. The resulting variables take the value 1 above the median and 0 otherwise. Measures of academic performance, such as percentile rank on the DNB examination, are converted into quartiles.

core subjects: physics & chemistry, biology, philosophy and modern language 1 and 2. Figure 5 displays the leave-one-out GTV distributions for these subjects. The *leave-one-out* GTV distribution for humanities-related subjects is shifted to the left compared with science-related subjects.[18] This suggests that teachers in philosophy and modern languages are less likely to use a gender-specific vocabulary in their feedback than math, physics and chemistry teachers, while biology teachers are somewhere in-between.[19] Philosophy is a particularly interesting point of comparison, since the gender composition of teachers in philosophy is quite similar to that in math-intensive subjects (62 percent of male teachers, see Appendix Table C4). Yet philosophy teachers seem to use a more gender-neutral vocabulary. This suggests that, for a given class, science teachers tend to differentiate their vocabulary more and confirms that our measure captures differences that go beyond characteristics specific to a class.

[Insert Figure 5 about here]

Finally, the fact that teachers' GTV is computed for multiple years and classes makes it possible to measure persistence of GTV across classes and over time. The correlation between a teacher's GTV and the *leave-one-out* GTV, i.e., the correlation between a given GTV and its average computed in other years×classes, is 0.092 and is significantly different from zero.[20] Note, however, that this correlation suffers from an attenuation bias, because we are correlating several GTVs measured with error, owing to the small sample used for the prediction at the class level. By way of comparison, in the teacher value-added literature, the within-teacher correlation is usually around 0.3 (Chetty et al., 2014). All in all, the GTV measure seems to capture persistent differences between the GTV of different teachers.

7.2 The Effect on Student Performance and Enrollment

Academic Performance. Panel A of Table 6 reports the estimated effect of having a teacher with a 1-standard-deviation higher *leave-one-out* GTV on students' standardized math performance on the *baccalauréat*, based on Equation (6). On average, being exposed to a teacher with a 1-standard-deviation higher GTV improves math performance on the *baccalauréat* by 1.4 percent of a standard deviation on average, a value significant at the 1 percent level. This effect corresponds to moving from a teacher with an average GTV to one at the 84$^{\text{th}}$ percentile

[18] Density distributions are all statistically different from each other at the 1 percent level, as suggested by pairwise Kolmogorov-Smirnov tests for equality (results available upon request).

[19] The median leave-one-out GTVs are as follows: 63.4 percent in physics and chemistry, 61.4 percent in biology, 61.4 percent in philosophy, 58.7 percent in modern language 1 and 59.6 percent in modern language 2.

[20] This correlation is obtained by regressing GTV on leave-one-out GTV. The significance we refer to in the text tests for whether the regression coefficient is statistically different from zero.

of the GTV distribution. The effect is slightly larger for female students, whose math grade increases by 1.8 percent of a standard deviation, than for male students (0.95 percent). The difference between the effects for male and female students is statistically significant at the 5 percent level.[21]

[Insert Table 6 about here]

Heteregeneity analysis for the effect on academic performance. The moderate average effects conceal a heterogeneity of responses depending on the degree of gender differentiation in the math teacher's feedback notes. Rather than include teacher GTV in the equation linearly, we explore the intensity of the treatment by regressing students' math grade on a set of GTV deciles, the first decile corresponding to the bottom 10 percent of the leave-one-out GTV distribution. Figure 6 plots the coefficients associated with the GTV deciles along with their 95 percent confidence intervals, separately for boys and girls. Compared with students exposed to the bottom 10 percent of teachers in terms of GTV, those with teachers in the 4$^{\text{th}}$ decile or above perform better on the *baccalauréat* exam by a significant 4 to 7.5 percent of a standard deviation on average, for both males and females. By contrast, we find no evidence of significant heterogeneity according to students' previous math performance or socio-economic status (see Appendix F section F.3 for details).

[Insert Figure 6 about here]

The last dimension of heterogeneity we explore is whether the type of vocabulary used by the teacher, i.e., the words more likely to be used for girls or boys, triggers different responses from students. While the general female-specific and male-specific vocabulary can be described and classified in broad categories, as in Section 6, this cannot be done at the teacher level, so as to properly distinguish different feedback styles (e.g. "encouraging", "competence-related"). We seek to disentangle the effect of exposure to a teacher who is relatively more likely to use the "female vocabulary" (or the "male vocabulary") for a given level of gendered feedback.

[Insert Table 7 about here]

To measure whether the positive effect on math performance at *baccalauréat* is reinforced by teachers more likely to use the vocabulary associated with female or with male students, we estimate Equation (6) augmented with GTV-male (or GTV-female) in addition to the

[21] We performed the following Wald test : $(0.0179 - 0.0095)/\sqrt{0.0031^2 + 0.0029^2} = 1.98$. The associated p-value is 0.048.

overall GTV measure (Table 7). The positive effect on math performance documented for higher-GTV teachers is reinforced for teachers with a higher GTV-female (those more likely to use the vocabulary associated with females). For a given level of GTV, a 1-standard-deviation increase in GTV-female is associated with an average additional increase in math performance of 0.65 percent of a standard deviation. Our results suggest that exposure to a higher GTV-female teacher matters only for girls, for whom the standardized math grade increases by an additional 0.9 percent of a standard deviation against a non-significant 0.5 percent for boys. Note however that the difference between the two coefficients is not statistically significant.[22] On the other hand, exposure to higher GTV-male teachers diminishes the positive effect on math performance and seems to be detrimental to female students, whose math performance worsens by 0.7 percent of a standard deviation, while male students are largely unaffected.[23] These results are consistent with both *potential mechanisms 2* and *6* discussed in Section 3 stating that i) feedback is more likely to enhance performance when it emphasizes efforts, and ii) male and female students react differently to feedback. They also echo results from the literature (see Huillery et al. (2024) for instance) who show that female students benefit more from growth mindset feedback compared to male students.

University applications and enrollment. Although having a math teacher with 1-standard-deviation higher GTV (moderately) improves female and male students' performance, we find no evidence of significant effects on their university application and enrollment outcomes. The effects on the probability of students' taking a STEM program as top choice in their applications (Table 6, Panel B) and of enrollment in a STEM undergraduate program (Table 6, Panel C) are small and mostly statistically insignificant at conventional levels. If anything, male students are marginally less likely (-0.3 percentage point) to top rank and enroll in a selective STEM program, a 2.2 percent decrease from the baseline probability of 21 percent. We find no evidence of any heterogeneous effects by decile of teacher GTV, by previous math performance or by socio-economic status. These results echo *potential mechanism 5* in which we hypothesized that the last ordinary feedback is unlikely to have long lasting effects on beliefs and, consequently, to change higher education choices and enrolment.

Robustness checks. Our results are robust to a series of sensitivity tests (data reported in Appendix Table F9). First, we check that our results are robust to controlling for students'

[22]We performed the following Wald test : $(0.0084 - 0.0051)/\sqrt{0.0038^2 + 0.0035^2} = 0.64$. The associated p-value is 0.52.

[23]Again, the difference between the two coefficients is not statistically different. Wald test : $(-0.0072 + 0.0048)/\sqrt{0.0034^2 + 0.0031^2} = -0.52$. The associated p-value is 0.60.

baseline characteristics (columns 1 and 4) and for the share of girls in the class (columns 2 and 5). Second, to strengthen the argument that the effect we estimate is specific to math teacher GTV, we control for the average GTV in the same class for the five other core subjects (Columns 3 and 6). Including these controls does not alter the magnitude or the significance of the results. The effect on math performance of a 1-standard-deviation increase in GTV ranges from 0.94 to 1.2 percent of a standard deviation for boys, and 1.6 to 1.8 percent for girls. The limited but significant effects on the probability of top-ranking a STEM program in the ROL and on the probability of enrolling in a STEM program (a decrease of about 0.3 percentage point) persist among boys. We also show in Appendix Table F10 that results are qualitatively similar when we use the GTV measure obtained from a random forest model. Finally, as placebo tests, Appendix Table F8 reports the effect of math teacher GTV on the standardized grades in physics, biology and philosophy. The effects are not statistically significant for any of these core subjects (except for the grade in physics for boys, significant at the 10 percent level), which is consistent with the idea that our baseline estimates capture the effect of the math teacher and not some unobserved differences between classes.

8 What is the Profile of Teachers with a Higher GTV ?

As our setting does not involve feedback manipulation, but only manipulation in exposure to teachers with different levels of gendered feedback, our GTV measure could be correlated with other teacher characteristics known to influence students' performance. The estimated coefficients, that is, might capture the effects of the latter. In this section, we consider two such characteristics : i) teachers' gender grading bias, with the hypothesis in mind that teachers using a gendered vocabulary could also encourage girls by overgrading them relative to boys ; ii) teachers' feedback personalization, investigating whether teachers using gendered feedback are also more likely to personalize their feedback notes. Section 8.1 describes the characteristics associated with teacher GTV, and section 8.2 reports the results of GTV on math performance accounting for each of those characteristics respectively.[24]

8.1 Teacher GTV and Other Teacher Characteristics

We compute two additional measures describing teaching styles: gender grading bias (Lavy and Sand (2018); Terrier (2020), see Appendix E for details) and a measure of feedback personalization.

[24]Given the small and mostly non-significant effects of GTV on higher education choices and enrollment outcomes, we show only the analysis of the mechanisms for the standardized grades in math. The results for other outcomes are available upon request.

We proxy the latter by computing a measure of within-teacher text distance. The greater the text distance between the different feedback notes, the more the teacher personalizes feedback.[25] Appendix Table E7 reports the results of a regression of the *leave-one-out* GTV on grading bias and feedback personalization, along with a dummy indicating whether the teacher is a male and whether he or she is a head teacher. It shows that a 1-standard-deviation increase in grading bias (i.e. a reduction in the grading bias favoring girls) is associated with a 0.04-standard deviation decrease in GTV, significant at the 1 percent level. Put differently, this correlation (even though rather moderate) suggests that teachers who use a more highly gendered vocabulary may also be slightly more likely to encourage female students through higher continuous assessment grades. Higher GTV is also positively correlated with feedback personalization (0.09 standard deviation increase), thus showing that teachers writing personalized notes are also more likely to use a gendered vocabulary.

8.2 The Effect of GTV on Performance Controlling for Teacher Observables

We reestimate equation 6 controlling for each teacher characteristic in turn to investigate the extent to which the effect of GTV on math performance is mediated by these characteristics. Controlling for the grading bias or feedback personalization does not affect the magnitude of the GTV effect. Panel (a) of Table 8 reports the estimated coefficients on GTV and on teachers' grading bias. A 1-standard-deviation increase in grading bias increases math performance by 1.3 percent of a standard deviation for boys and girls, a magnitude comparable to our effects of exposure to gendered feedback. But, its inclusion as a control does not alter the estimated coefficients on GTV, so we can rule out grading bias as a first-order mediator of the impact of GTV. While exposure to a teacher with personalized feedback appears to be beneficial to both boys and girls, having a teacher whose feedback is gendered still improves math performance significantly, by 0.9 percent of a standard deviation for boys and 1.7 percent for girls (Panel b of Table 8). Therefore, we can also discard the thesis that having a teacher with higher GTV affects performance through feedback personalization rather than through gender differentiation.

[Insert Table 8 about here]

Altogether, the additional results of this section suggest that exposure to gendered feedback

[25] More specifically, each feedback note is transformed into a vector of words. The Euclidean distance is computed for each pair of word vectors, and then averaged. In order to capture personalization that neutralizes the use of a gendered vocabulary, the measure of text distance is computed separately on the word vectors of male and female students.

affects students' performance, over and above other teacher practices such as grading or feedback personalization. This suggests that a (relatively small) component of teacher effectiveness is linked to the use of gendered language.

9 Conclusion

Using comprehensive administrative data on about 600,000 Grade 12 students' transcripts, this paper shows how the student's gender affects the vocabulary used by math teachers in their written feedback. To identify gendered patterns in teacher feedback notes, we apply machine learning techniques to predict students' gender from the words that the teachers elect. The key findings are threefold. First, we find that equally able female and male students get different feedback, and that the scope of this gender differentiation is ample: overall, the words used by Grade 12 math teachers allow correct prediction of the sex of 64 percent of the students. For comparison, this is only marginally lower than the feedback's predictive power as regards student performance, which is the upper bound of what we could expect given that the express purpose of the feedback is to assess performance. Second, qualitative analysis of the best gender predictors reveals that for female students, teachers tend to emphasize positive behavioral aspects and to encourage effort, while equally performing males are both more severely criticized for unruly behavior and more praised for intellectual skills.

We take the analysis one step further by investigating how gendered feedback relates to student performance, college applications and enrollment decisions in the following year. We compute a teacher-level measure of gendered teacher vocabulary (GTV), which we define as the share of students whose gender is correctly predicted. Exploiting the quasi-random assignment of teachers within high schools conditional on the elective courses chosen by students, we relate GTV to educational outcomes. The third key finding is that exposure to a teacher with higher GTV improves math performance by between 1.5 percent and 7.5 percent of a standard deviation, with slightly larger effects for girls than for boys. Seeking to get inside the black box of gendered feedback to understand which aspects of feedback potentially may help to determine student performance, we compare the effects of exposure to teachers who make greater or lesser use of the vocabulary associated with female or male students, for a given level of gendered feedback. We find that the effect of gendered feedback on boys' math performance does not vary with the type of vocabulary, but for girls it is greater when the teacher uses the vocabulary associated with female students, i.e. when teachers underscore the girls' positive behavior and effort, and detrimental when the teacher makes greater use of the male-type vocabulary. We find no significant or sizeable effects on higher education applications or enrollment decisions in the subsequent year.

The main take-away from our study should be an awareness message. Our findings indicate that the type of vocabulary used in teachers' written feedback may be an effective pedagogical device to improve students' performance but may require some tailoring given that male and female students react differently depending on whether, and how, their efforts or skills are praised. The paper implicitly suggests avenues for future research. Since our setting does not involve feedback manipulation, but only manipulation in exposure to teachers with different levels of gendered feedback, we cannot properly identify the features of gendered vocabulary that trigger the strongest impact on performance. To go a step further, we need additional research using experimental settings in which teacher feedback types vary randomly.

References

Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz, "What we teach about race and gender: Representation in images and text of children's books," *The Quarterly Journal of Economics*, 2023, *138* (4), 2225–2285.

Alan, S., T Boneva, and S. Ertac, "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit," *Quarterly Journal of Economics*, 2019, *134* (3), 1121–1162.

Andersen, Ida Gran, "Teachers' gender bias in STEM: Results from a vignette study," *British Educational Research Journal*, 2023, *49* (4), 833–851.

Bassi, M., M. Díaz, R. Blumberg, and A. Reynoso, "Failing to notice? Uneven teachers' attention to boys and girls in the classroom," *IZA Journal of Labor Economics*, 2018, *7* (1), 1–22.

Blackwell, L., K. Trzesniewski, and C. Dweck, "Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention," *Child Development*, 2007, *78*, 246–263.

Bobba, M. and V. Frisancho, "Self-Perceptions about Academic Achievement: Evidence from Mexico City," *Journal of Econometrics*, 2020.

Borhen, A., A. Imas, and M. Rosenberg, "The Language of Discrimination: Using Experimental versus Observational Data," *AEA Papers and Proceedings*, 2018, *108*, 169–174.

Breda, Thomas, Julien Grenet, Marion Monnet, and Clémentine Van Effenterre, "How Effective are Female Role Models in Steering Girls Towards STEM? Evidence from French High Schools," *The Economic Journal*, 2023, *133* (653), 1773–1809.

Burnett, P., "Teacher Praise and Feedback and Students' Perceptions of the Classroom Environment," *Educational Psychology*, 2002, *22* (1).

Burnette, Jeni L, Joseph Billingsley, George C Banks, Laura E Knouse, Crystal L Hoyt, Jeffrey M Pollack, and Stefanie Simon, "A systematic review and meta-analysis of growth mindset interventions: For whom, how, and why might such interventions work?," *Psychological bulletin*, 2023, *149* (3-4), 174.

Canning, E., E. Ozier, H. Williams, R. AlRasheed, and M. Murphy, "Professors Who Signal a Fixed Mindset About Ability Undermine Women's Performance in STEM," *Social Psychological and Personality Science*, 2021.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri, "Gender Differences in Peer Recognition by Economists," *NBER Working Paper*, 2021, *28942.*

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

Chetty, R., J. Friedman, and J. Rockoff, "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 2014, *104* (9), 2593–2632.

Cnesco, *La constitution des classes: pratiques et enjeux*, Paris: Cnesco, 2015.

Coffman, Katherine, Maria Paola Ugalde Araya, and Basit Zafar, "A (dynamic) investigation of stereotypes, belief-updating, and behavior," *Economic Inquiry*, 2024.

Corpus, J. and M. Lepper, "The Effects of Person Versus Performance Praise on Children's Motivation: Gender and Age as Moderating Factors," *Educational Psychology*, 2007, *27* (4), 487–508.

Delavande, A., E. Del Bono, A. Holford, and S. Sonkurt, "Skills Accumulation with Malleable Ability: Evidence from a Growth Mindset Intervention," *Working Paper*, 2019.

Dupas, P., A. Handlan, A. Modestino, M. Niederle, M. Sere, H. Sheng, and J. Wolfers, "Gender Differences in Economics Seminars," *American Economic Review*, 2025.

Dweck, C., *Mindset: The New Psychology Of Success*, New York: Random House, 2006.

Franco, C., "How Does Relative Performance Feedback Affect Beliefs and Academic Decisions?," *Working Paper*, 2019.

Gale, David E. and Lloyd S. Shapley, "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 1962, *69* (1), 9–15.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as data," *Journal of Economic Literature*, 2019, *57* (3), 535–74.

Hakimov, R., R. Schmacker, and C. Terrier, "Confidence and College Applications: Evidence from a Randomized Intervention," *Working Paper*, 2023.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media, 2009.

Huillery, E., A. Bouguen, A. Charpentier, Y. Algan, and C. Chevallier, "The Role of Mindset in Education: A Large-Scale Field Experiment in Disadvantaged Schools," *Economic Journal*, 2024.

Koffi, M., "Innovative Ideas and Gender (In)equality," *American Economic Review*, 2025.

Landaud, Fanny, Éric Maurin, Barton Willage, and Alexander Willén, "The value of a high school gpa," *Review of Economics and Statistics*, 2024, pp. 1–24.

Lavy, Victor and Edith Sand, "On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases," *Journal of Public Economics*, 2018, *167* (C), 263–279.

Lovász, Anna, Ewa Cukrowska-Torzewska, Mariann Rigó, and Ágnes Szabó-Morvai, "Gender differences in the effect of subjective feedback in an online game," *Journal of Behavioral and Experimental Economics*, 2022, *98*, 101854.

Ly, Son Thierry and Arnaud Riegert, "Mixité sociale et scolaire et ségrégation inter—et intra-établissement dans les collèges et lycées français," *Rapport du Conseil national d'évaluation du système scolaire (CNESCO). Téléacessible à: http://www. cnesco. fr/wp-content/uploads/2015/05/Etat-des-lieux-Mixité-à-lécoleFrance1. pdf*, 2015.

Macnamara, B. and A. Burgoyne, "Do growth mindset interventions impact students' academic achievement? A systematic review and meta-analysis with recommendations for best practices.," *Psychological bulletin*, 2023, *149* (3-4), 133.

Morgan, C., "The Effect of Negative Managerial Feedback on Student Motivation: Implications for Gender Differences in Teacher-Student Relations," *Sex Roles*, 2001, *44.*

Ningrum, P., T. Pansombut, and A. Ueranantasun, "Text Mining of Online Job Advertisements to Identify Direct Discrimination During Job Hunting Process: A Case Study in Indonesia," *PLoS ONE*, 2020, *15* (6).

Owen, S., "College Field Specialization and Beliefs about Relative Performance," *Working Paper*, 2022.

Roth, Alvin E, "The economics of matching: Stability and incentives," *Mathematics of operations research*, 1982, *7* (4), 617–628.

Sarsons, H., C. Gerxhani, E. Reuben, and A. Schram, "Gender Differences in Recognition for Group Work," *Journal of Political Economy*, 2021, *129* (1).

Sarsons, Heather, "Interpreting Signals in the Labor Market: Evidence from Medical Referrals," *Working Paper*, 2019.

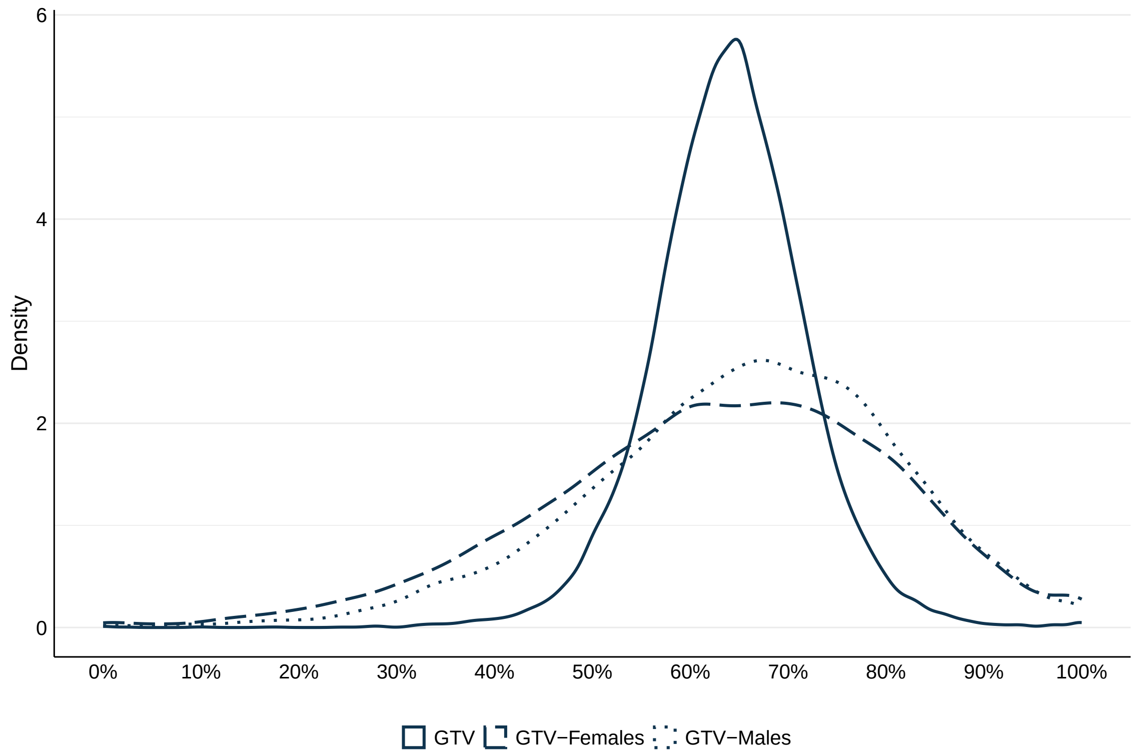
Terrier, C., "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement," *Economics of Education Review*, 2020, *77.*

Wu, A., "Gendered Language on the Economics Job Market Rumors Forum," *AEA Papers and Proceedings*, 2018, *108*, 175–179.

Wu, X., C. Li, Y. Zhu, and Y. Miao, "Short Text Topic Modeling with Topic Distribution Quantization and Negative Sampling Decoder," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

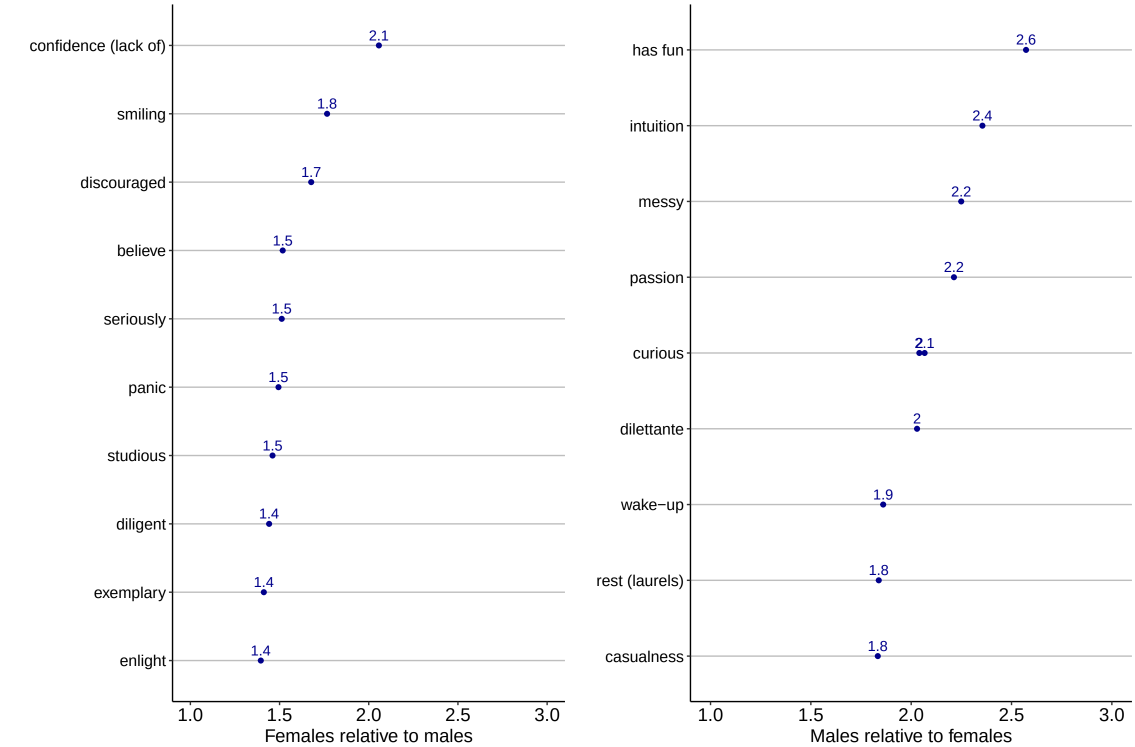
Zimmermann, F, "The Dynamics of Motivated Beliefs," *American Economic Review*, 2020, *110* (2), 337–363.

Figure 1 – Distribution of Math Teachers' Overall *Leave-one-out* GTV and by Students' Gender



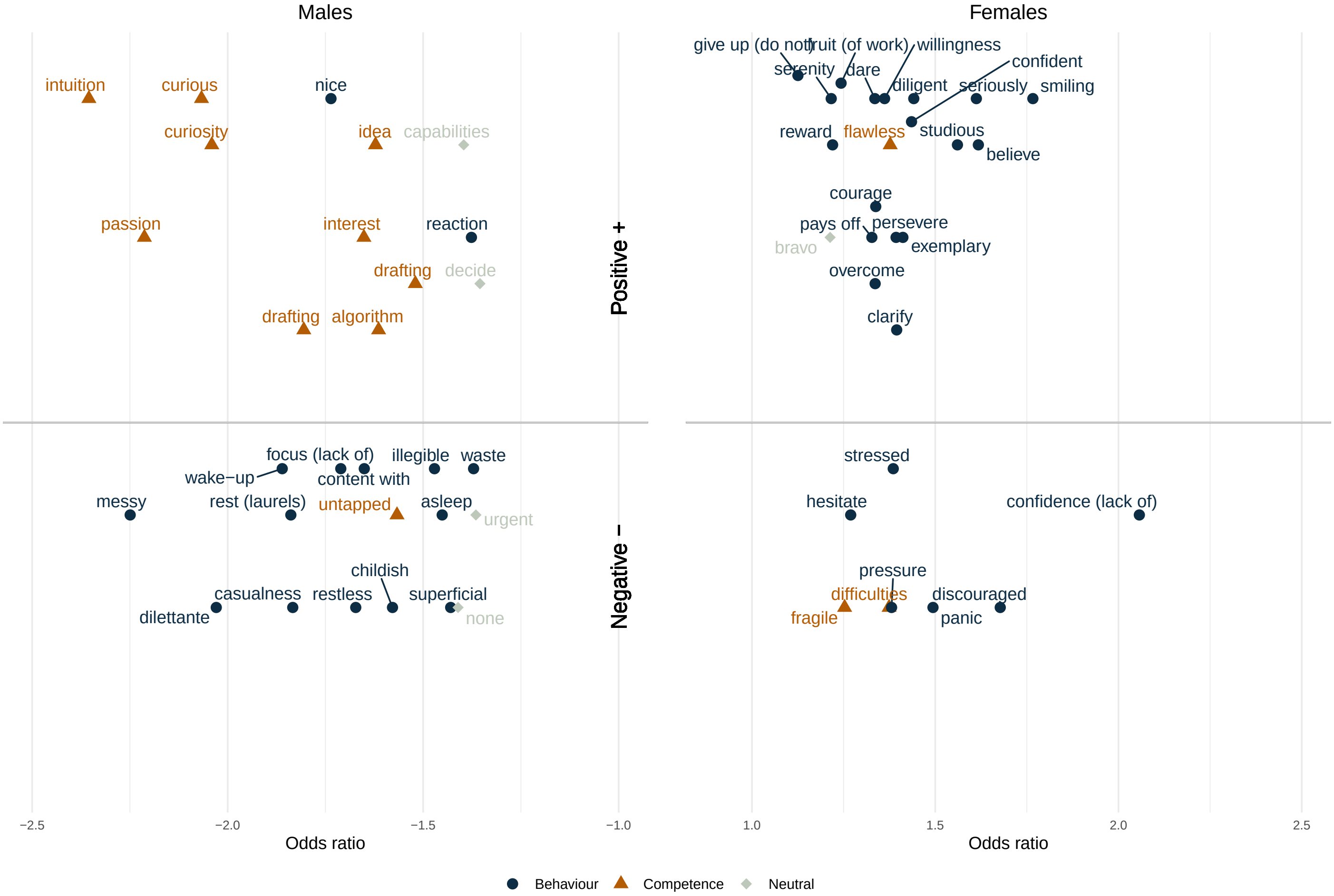
Notes: This figure shows the distributions of math teachers' overall *leave-one-out* GTV, as well as the teacher accuracy computed for female students (*leave-one-out* GTV-females) and male students respectively (*leave-one-out* GTV-males).

Figure 2 – Odds Ratios of the Top 10 Gender Predictors



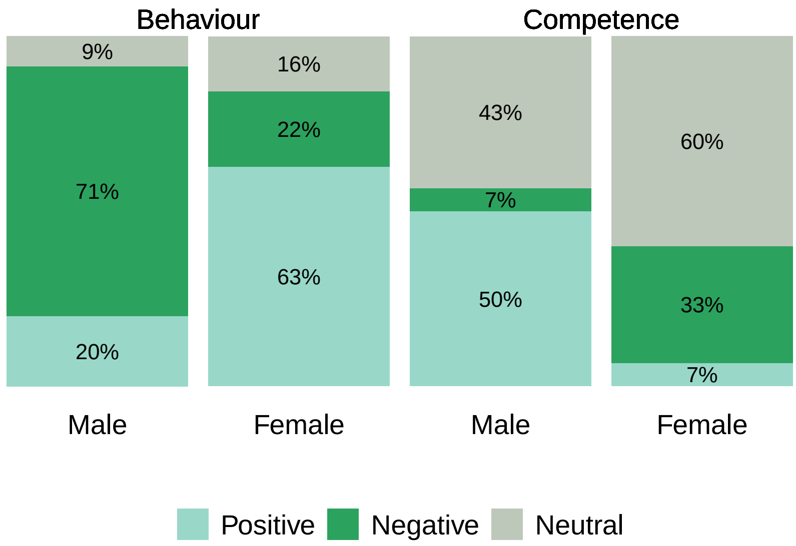
Notes: This figure shows the odds ratios obtained for the 10 best predictors of the female gender (left-hand side), and for the 10 best predictors of the male gender (right-hand side). These odds ratios are obtained from the estimation of the model described by Equation (1), where the vocabulary appearing in math teachers' feedback was used to predict student gender.

Figure 3 – Classification of the Top 30 Gender Predictors

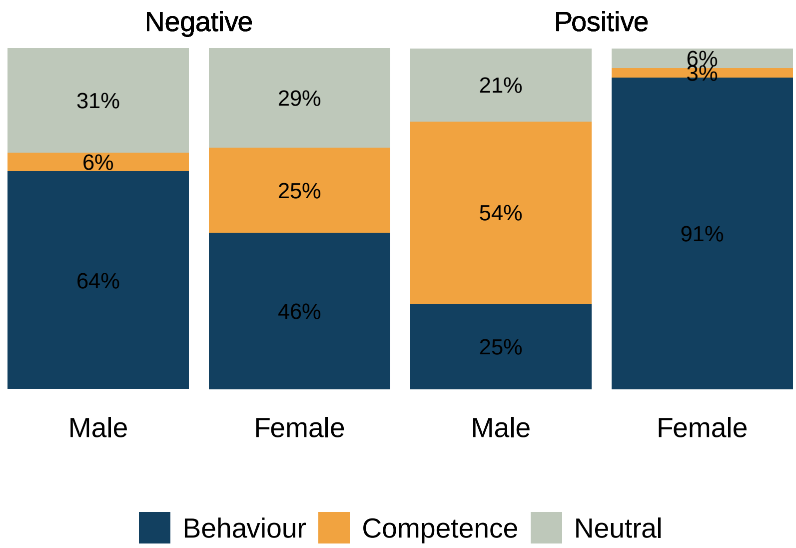


Notes: This figure classifies the top 30 female and top 30 male predictors of the model described by Equation (1) estimated using the vocabulary appearing in math teachers' feedback into positive vs. negative and behavioural vs. competence-related categories. Ambiguous words (i.e. the ones used in both positive and negative contexts) or words that do not fit in any of the categories are respectively labelled neutral or unclassified. The x-axis gives the odds-ratio of each predictor.

Figure 4 – Gender Predictors' Type and Positiveness



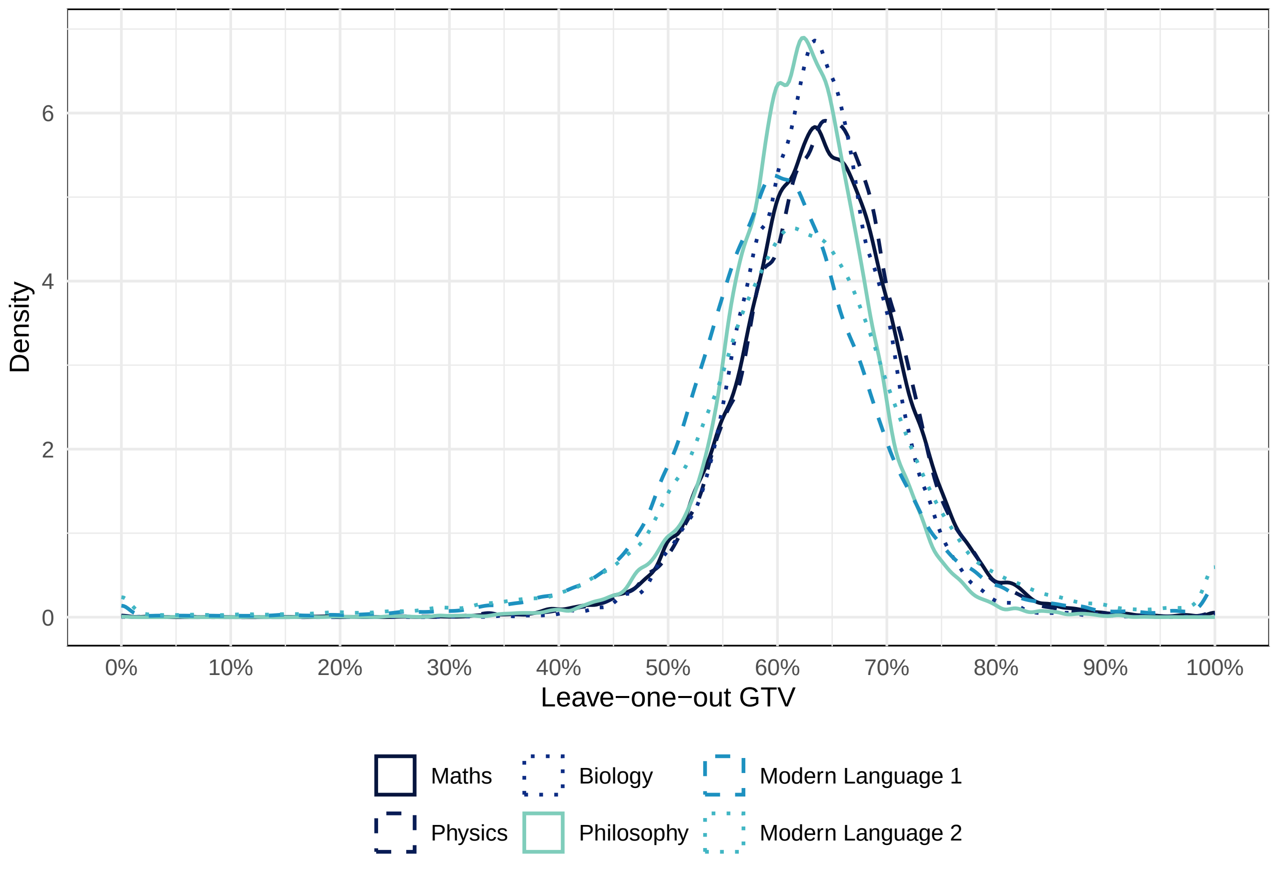
(a) Positiveness conditional on feedback type



(b) Feedback type conditional on positiveness

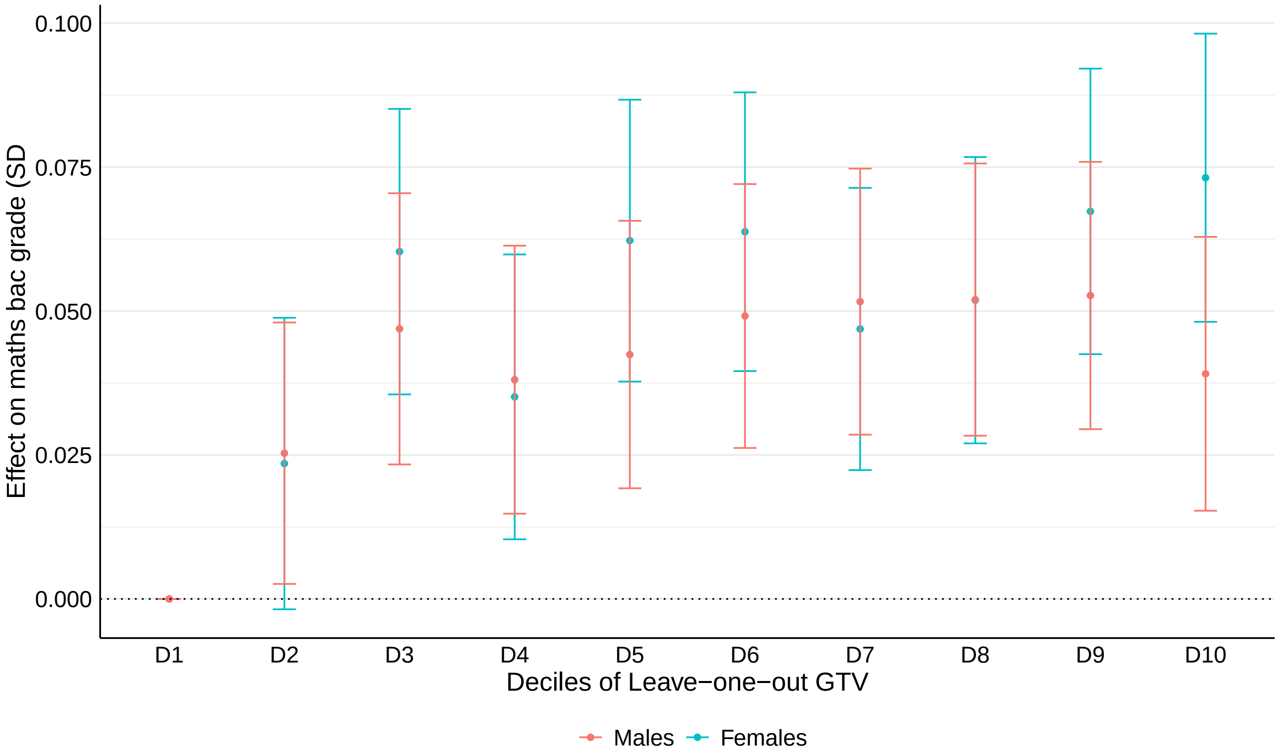
Notes: This figure used the classification of the top 100 male and top 100 female predictors obtained by the estimation of the model described by Equation (1), where the vocabulary appearing in math teachers' feedback was used to predict student gender. The predictors are classified into positive vs. negative and behavioural vs. competence-related categories. Ambiguous words (i.e., the ones used in both positive and negative contexts) or words that do not fit in any of the categories are respectively labelled neutral or unclassified. Panel a shows the proportions of positive, neutral and negative gender predictors conditional on being competence-related or behavioural, and Panel b shows the proportions of behavioural and competence-related gender predictors conditional on positiveness.

Figure 5 – Distribution of Teachers' *Leave-one-out* GTV – By Core Subjects



Notes: This figure shows the distributions of the math, physics, biology, philosophy and foreign language teachers' leave-one-out GTV measure. Density distributions are all statistically different from each other at the 1 percent level as suggested by pairwise Kolmogorov-Smirnov tests for equality.

Figure 6 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Math Performance - By GTV Deciles



Notes: The figure reports the results of the regression of students' standardised grade on the math *baccalauréat* exam on a set of teacher *leave-one-out* GTV decile dummies, controlling for high school, year and elective fixed effects. Coefficients are expressed in deviation from the first decile's value, and are reported with their 95% confidence intervals.

Table 1 – Number of Grade 12 Science Major Students and Sample Restrictions

	2013	**2014**	**2015**	**2016**	**2017**
Total nb. of G12 science major students	179,625	183,693	190,980	198,573	203,262
Nb. of obs. with missing transcript	81,515	58,624	46,973	38,929	22,946
% high school entirely missing:	90	84	77	68	84
High schools less 2 classes	15,649	15,573	15,219	15,335	15,186
Teachers less 2 classes	6,047	6,363	6,475	9,762	35,259
Obs. in sample	**87,808**	**112,944**	**131,470**	**143,363**	**138,883**
in %	**48.88**	**61.49**	**68.84**	**72.20**	**68.33**

Notes: This table reports the number of Grade 12 science major applicants on the APB platform for each year. We show the number of observations removed for each sample restriction, and provide the number of observations used in the analytical sample in bold in the table. "High school entirely missing" refers to students enrolled in high schools that do not report grade transcripts automatically on the APB platform and that are therefore discarded from the sample.

Table 2 – Grade 12 Science Major Students' Summary Statistics

	All	**Males**	**Females**
Demographics			
Female	0.47		
Age (years)	18.08	18.10	18.05
Free lunch student	0.13	0.11	0.14
High SES	0.43	0.45	0.42
Medium-high SES	0.16	0.16	0.16
Medium-low SES	0.24	0.23	0.25
Low SES	0.17	0.16	0.18
Education: past academic performance			
Rank at DNB: French	0.49	0.43	0.55
Rank at DNB: math	0.49	0.50	0.47
Rank at baccalaureat French (written)	0.46	0.41	0.52
Rank at baccalaureat : French (oral)	0.46	0.42	0.51
Education: G12 elective course choice			
Maths	0.24	0.27	0.19
Physics-chemistry	0.26	0.27	0.25
Earth and life science	0.37	0.26	0.49
Engineering and computer science	0.13	0.20	0.06
Nb. of Observations	614,466	325,653	288,813

Notes: This table shows descriptive statistics for Grade 12 science major students on the whole analytical sample, and separately for boys and girls. The number of observations reported in the last row is the number of observations for which all the observable characteristics are non-missing.

Table 3 – Math Teachers' Summary Statistics

	Mean	**S.d**
Share of head teacher at least once	0.55	0.50
Male math teacher	0.57	0.49
Number of teacher observations	3.95	1.61
Average number of classes per year	1.12	0.30
Average number of students per prof/year	27.71	4.95
Average feedback length	14.89	4.92
Nb. of teachers	6,716	

Notes: This table shows descriptive statistics for math teachers in the analytical sample teaching Grade 12 Science Major students. The average feedback length is computed as the average number of words in teachers' feedback, once common words (such as *the*, *she*, *a*, etc.) have been removed.

Table 4 – Balancing Test: Leave-One-Out GTV with Students' Baseline Characteristics

	Dep. var: Teacher's leave-one-out GTV (SD)		
Variable	**Coeff.**	**S.e**	**p-value**
Female	-0.0076	0.0303	0.8019
Age (years)	-0.0201	0.0254	0.4292
Foreign student	0.0541	0.0892	0.5443
Free lunch student	0.0675	0.0464	0.1455
High SES	0.0307	0.0237	0.1943
Medium-high SES	-0.0296	0.0311	0.3413
Medium-low SES	-0.0344	0.0237	0.1476
Low SES	0.0413	0.0293	0.1597
Rank at DNB: math	-0.0513	0.0772	0.5065
Rank at DNB: French	-0.0082	0.0628	0.8955
Rank at baccalaureat : French (written)	0.0868	0.0580	0.1343
Rank at baccalaureat : French (oral)	-0.0567	0.0571	0.3207
High school x elective x year FE	Yes		
Nb. of observations	58,359		

Notes: This table reports the estimation results of the standardized *leave-one-out* GTV measure, defined at the class-level, regressed on students' socio-economic characteristics and baseline academic performance. The regression includes high school×elective course× year fixed effects. Standard errors are clustered at the teacher level and are reported in the second column. ***: p-value < 0.01; **: p-value < 0.05, *: p-value < 0.1.

Table 5 – Pearson's Chi Square Tests of Class Random Assignement

Variable	**N non-miss pvalues**	**N sig pvalues at 5%**	**Sh. sig at 5%**	**Sh. sig at 1%**
Female	19664	1487	7.56	2.16
Age (years)	19651	1093	5.56	1.89
Foreign student	6963	194	2.79	1.09
Free lunch student	16916	545	3.22	0.95
High SES	19854	844	4.25	0.94
Medium-high SES	18610	527	2.83	0.58
Medium-low SES	19590	622	3.18	0.61
Low SES	18095	624	3.45	0.76
Rank at DNB: math	18918	111	0.59	0.10
Rank at DNB: French	18921	79	0.42	0.06
Rank at baccalaureat : French (written)	20329	580	2.85	0.69
Rank at baccalaureat : French (oral)	20332	587	2.89	0.66

Notes: This table reports the results of the Pearson Chi-square tests of independance performed on the unique combinations of high schools, elective course and year. For each unique combination, we tabulate math teachers' identifiers with each baseline characteristic. Continuous variables such as age and percentile ranks are first discretized. Columns 3 and 4 report the share of p-values that are above the nominal levels of 5 percent and 1 percent respectively.

Table 6 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Educational Outcomes

	All	**Boys**	**Girls**
Academic performance			
Grade at baccalauréat (SD): math	0.0135***	0.0095***	0.0179***
	(0.0025)	(0.0029)	(0.0031)
Type of program ranked first in the ROL			
All Sciences	-0.0018+	-0.0034*	0.0007
	(0.0010)	(0.0014)	(0.0013)
Selective Sciences	-0.0013	-0.0033**	0.0010
	(0.0009)	(0.0013)	(0.0010)
University Sciences	-0.0005	-0.0010	0.0002
	(0.0005)	(0.0007)	(0.0007)
Vocational Sciences	-0.0001	0.0010	-0.0008
	(0.0006)	(0.0010)	(0.0007)
Matriculation in the following year			
All Sciences	-0.0017+	-0.0031*	-0.0005
	(0.0009)	(0.0013)	(0.0012)
Selective Sciences	-0.0013*	-0.0022*	-0.0002
	(0.0006)	(0.0009)	(0.0007)
University Sciences	-0.0016*	-0.0019	-0.0017
	(0.0008)	(0.0012)	(0.0011)
Vocational Sciences	-0.0003	-0.0011	0.0006
	(0.0006)	(0.0010)	(0.0006)
Nb. of observations	609331	320996	285884

Notes: This table reports the coefficients on the standardized *leave-one-out* GTV obtained from the estimation of Equation (6). Each row corresponds to a different linear regression performed on the full analytical sample and separately by gender, with the dependent variable listed on the left. The regression includes high school×elective course×year fixed effects. The number of observations is lower than in the analytical sample because singletons—high school×year×elective cells with the same teacher (and thus identical GTV)—do not contribute to coefficient identification. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.001; **: p-value < 0.01, *: p-value < 0.5, +: p-value < 0.1.

Table 7: Effects of Exposure to Male- of Female-Specific Vocabulary on Students' Math Performance

	All	**Boys**	**Girls**
Panel a. Higher exposure to female vocabulary			
Grade at baccalaureat (SD): math			
GTV one-out	0.0097**	0.0062+	0.0132***
	(0.0030)	(0.0035)	(0.0037)
GTV one-out female	0.0064*	0.0051	0.0084*
	(0.0031)	(0.0035)	(0.0038)
Panel b. Higher exposure to male vocabulary			
Grade at baccalaureat (SD): math			
GTV one-out	0.0162***	0.0116***	0.0214***
	(0.0027)	(0.0031)	(0.0034)
GTV one-out male	-0.0057*	-0.0048	-0.0072*
	(0.0027)	(0.0031)	(0.0034)
Nb. of observations	601708	316742	282556

Notes: Each panel reports the coefficients on the standardized *leave-one-out* GTV obtained from the estimation of Equation (6) augmented with GTV-female (Panel a) or GTV-male (Panel b.), where the outcome is the standardized grade in math at the *baccalauréat*. It is estimated on the whole sample and separately for Grade 12 male and female students. The regression includes high school×elective course×year fixed effects. The number of observations is lower than in the analytical sample because singletons—high school×year×elective cells with the same teacher (and thus identical GTV)—do not contribute to coefficient identification. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.001; **: p-value < 0.01, *: p-value < 0.5, +: p-value < 0.1.

Table 8: Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Math Performance - Controlling for Teacher Characteristics

	All	**Boys**	**Girls**
Panel a. Teacher grading bias (SD)			
GTV one-out	0.0145***	0.0101***	0.0193***
	(0.0025)	(0.0029)	(0.0032)
Grading bias	0.0126***	0.0134***	0.0129***
	(0.0026)	(0.0030)	(0.0034)
Panel b. Teacher Feedback Personalization (SD)			
GTV one-out	0.0125***	0.0087**	0.0167***
	(0.0025)	(0.0029)	(0.0031)
Feedback Pers.	0.0152***	0.0109**	0.0184***
	(0.0031)	(0.0035)	(0.0038)
Nb. of observations	601972	316871	282693

Notes: This table reports the coefficients on the standardized *leave-one-out* GTV obtained from the estimation of Equation (6) (row *GTV one-out*). Each panel corresponds to a different linear regression performed on the full analytical sample and separately by gender, with the dependent variable being the standardized grade in math at the *baccaulauréat.* The regression further controls for the standardized teacher grading bias (Panel a.) and for the standardized measure of teacher feedback personalization (Panel b.). The regression includes high school×elective course×year fixed effects. The number of observations is lower than in the analytical sample because singletons—high school×year×elective cells with the same teacher (and thus identical GTV)—do not contribute to coefficient identification. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.001; **: p-value < 0.01, *: p-value < 0.5, +: p-value < 0.1.

Appendix to

Gendered Teacher Feedback, Students' Math Performance and Enrollment Outcomes: A Text Mining Approach

Pauline Charousset, Marion Monnet

September 2025

List of Appendices

A Gendered Teacher Vocabulary (GTV): Details of the Estimation Procedure and Additional Results

This appendix details the practical implementation of the steps taken in estimating the gendered teacher vocabulary (GTV) in Section 5.

A.1 Textual Data Preparation

The students' academic records consist of a corpus of *documents*, where a *document* corresponds to the feedback note a teacher wrote to a given student, in a given subject. Our aim is to convert all the documents into a data structure similar to the one displayed in Table A1. In this example, all the words and groupings of two words that appear at least once in a document have been converted to a column.

Text cleaning. To reduce the dimensionality of our data and, consequently, the computational burden of our estimation, we follow the text cleaning steps suggested by Gentzkow et al. (2019). For each *document*, we remove punctuation signs, but keep track of the position of full stops to identify the different sentences of the original text. We get rid of first names (identified with the Insee register of French first names), which would be very good predictors of student gender without reflecting any gender differentiation in the vocabulary used. We also remove *stop words*, which are very common words that bear little informational content, like "*le*" ("the"), "*donc*" ("thus"), "*déjà*" ("already"), etc.

All remaining words are *stemmed*, i.e. replaced by their roots: for instance, the words "*amateur*" and "*amatrice*" are replaced by their common root "*amat*". This last step is crucial to our analysis, because it allows to get rid of all the grammatical markers of the students' gender, which often appear, in French, at the end of the words. We further reduce the dimensionality of our data by getting rid of all *stemmed* words that appear in less than 100 documents.

Tokenization. To convert the remaining words into a set of columns (also known as the document-term matrix), we "dummify" words and grouping of words. Each word that appears in the corpus becomes a column, that takes value 1 if the word appears in the document, and 0 otherwise. In the text analysis literature, groups of words are commonly denoted *ngrams*, were n is the number of words in the considered group of words. In our analysis, we choose to use *unigrams*, i.e. one-word tokens, as regressors, and perform a robustness checks adding *bigrams* to the set of predictors.

Table A1 – From text to data: an illustration

Document	ensemble	alarmant	bon	travail	sérieux	ensemble alarmant	bon travail
Ensemble alarmant, manque de sérieux.	1	1	0	0	1	1	0
Bon travail, beaucoup de sérieux.	0	0	1	1	1	0	1

A.2 Predicting Student Gender and Measuring GTV

In this second step, the tokens are used as gender predictors. We assume that the probability of being a female student conditional on the words used in the feedback has a logistic form:

$$P(Female_i = 1|W_i) = \frac{exp(\alpha W_i)}{1 + exp(\alpha W_i)} \quad \forall i \tag{A.1}$$

and our objective is to find the set of α coefficients that minimize a penalized version of the negative log-likelihood $\ln(L(\alpha))$, where λ denotes the regularization parameter chosen via a cross-validation procedure:

$$\hat{\alpha} = argmax_{\alpha}(\ln(L(\alpha)) - \lambda \sum_{w=1}^{W_n} |\alpha_w|) \tag{A.2}$$

The $\hat{\alpha}$ estimates are then used to predict gender. The GTV measure, i.e. the proportion of a teacher's students for whom the model correctly predicts gender, is computed based on these predictions. In practice, we estimate a logistic Lasso to determine the $\hat{\alpha}$ coefficients. We detail below the practical implementation of the estimation.

Step 1: Undersampling. Before any estimation is done, we use undersampling techniques to construct an estimation sample such that no correlation subsists between gender and math performance. For each class, we sample as many male and female students at each quartile of prior math ability (proxied by DNB percentile rank in math). Then, for each class×quartile, we select n_{cq} males and n_{cq} females where $n_{cq} = min(n_{cq}^{females}; n_{cq}^{males})$.[A.1]

Step 2: Random selection of tokens. As shown in Table 3, the number of tokens used in feedback varies by teacher. As feedback length could influence the quality of the prediction, we randomly sample fifteen tokens for lengthy feedback, defined as the ones with an above-median length (15 words).

Step 3: Training and hold-out samples. To avoid overfitting concerns, we fit model A.2 on a training sample (30 percent of the undersampled data) and predict gender on a hold-out sample (70 percent). To preserve the balanced structure of the undersampled data, the partition of the data into a training and a hold-out sample is stratified, i.e. we include 30 percent (70 percent) of n_{cq} males and females in the training (hold-out) sample.

Step 4: Training the model. The training sample is used to fit the model and get the estimated $\hat{\alpha}$ coefficients. We first tune the regularization parameter λ by running a logistic Lasso with a 10-fold cross validation. We pick the λ value that lies within one standard deviation of the minimal error (Hastie et al., 2009) and estimate the logistic-lasso to obtain the $\hat{\alpha}$.

Step 5: Predict students' gender. The fitted model is applied to the hold-out sample to predict each student's gender. The model classifies a student as a girl ($\widehat{Sex}_i = 1$) if the predicted probability is greater than 0.5, and as boy otherwise ($\widehat{Sex}_i = 0$).

Step 6: Compute the GTV measure. Finally, for each class c of teacher j, we compute the GTV measure as the average proportion of correctly classified students:

[A.1] We use the French grade obtained on the DNB exam instead of the math grade when we compute the teacher GTV for humanities related subjects.

$$GTV_{jc} = \frac{1}{N_{jc}} \sum_{i=1}^{N_{jc}} \mathbb{1}\{Sex_i = \widehat{Sex_i}\} \times 100 \quad \forall j, c \tag{A.3}$$

where N_{jc} is the number of students in the balanced subsample of teacher j's students from class c:

$$N_{jc} = \sum_{c=1}^{C_j} \sum_{q=1}^{4} 2 \times n_{cq}$$

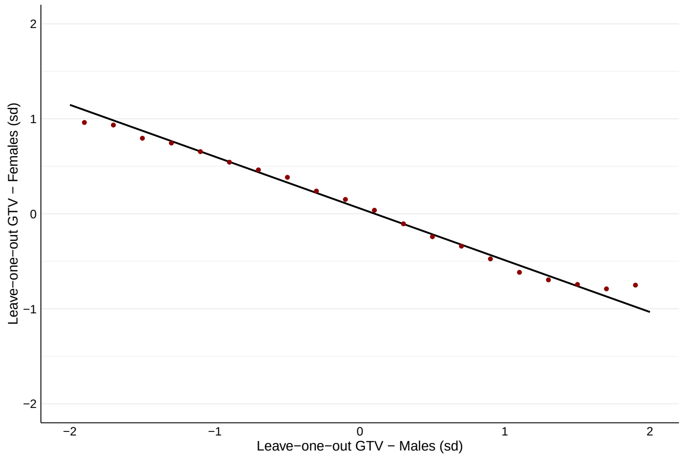
The GTV measure defined by Equation (A.3) could capture some unobserved-class specific gender differences. To allay this concern, we also compute the *leave-one-out* GTV as the average GTV over all the classes taught during the sample period, excluding class c:

$$GTV_{j\backslash c} = \frac{1}{N_j - 1} \sum_{c' \neq c} GTV_{jc'} \quad \forall j, c \tag{A.4}$$

The two GTV measures are inherently noisy as they are computed on a limited number of observations (N_{jc} is at most 102 in our sample). To stabilize those two measures and in order for our results not to depend on a single data split defined at Step 2, we repeat Step 1 to Step 5 100 times and use the GTV measures averaged over those 100 iterations.

A.3 Correlation Between the Leave-one-out GTV by Gender (GTV-Male and GTV-Female)

Figure A1 – Correlation of Leave-one-out GTV by Students' Gender



Notes: This figure plots the binned average of the *leave-one-out* GTV-female for different values of the *leave-one-out* GTV-male, along with the fitted regression line.

B Additional Results on Feedback Classification

B.1 Classification of the Top 100 Gender Predictors

Table B2 – Top 100 Predictors' Classification - Boys (in French)

	Positive	**Negative**	**Neutral**
Intellectual	aptitude, original, capable, finesse, ambition, potentiel, facilité, capacité, pertinent, idée, intérêt, curiosité, curieux, passion, intuition	lent, inexploité	scientifique, expression, domaine, écrit, copie, écrire, argument, rédige, algorithme, rédactionnel
Behavioral	rapidité, moteur, rigoureux, sympathique	néglige, dispersé, pénible, déborde, désordonné, inexistant, bâclé, survolé, nonchalant, gâché, superficiel, endormi, illisible, puéril, content, agité, canalise, désinvolture, repose, réveil, dilettante, brouillon, amuse	soin, réaction
Unclassified	conseil, mieux, confirme, meilleur, possibilité	manque, cruel, tard, refus, minimal, regret, dommage, perfectible, minimum, urgent, aucun	total, indispensable, absolu, étiez, impose, dure, maison, vitesse, sursaut, détaille, dépend, exige, attendre, passage, personnel, mesure, flagrant, agir, exploite, remettre, mur, mettre, suffisant, pris, vivre, mis, décide

Table B3 – Top 100 Predictors' Classification - Girls (in French)

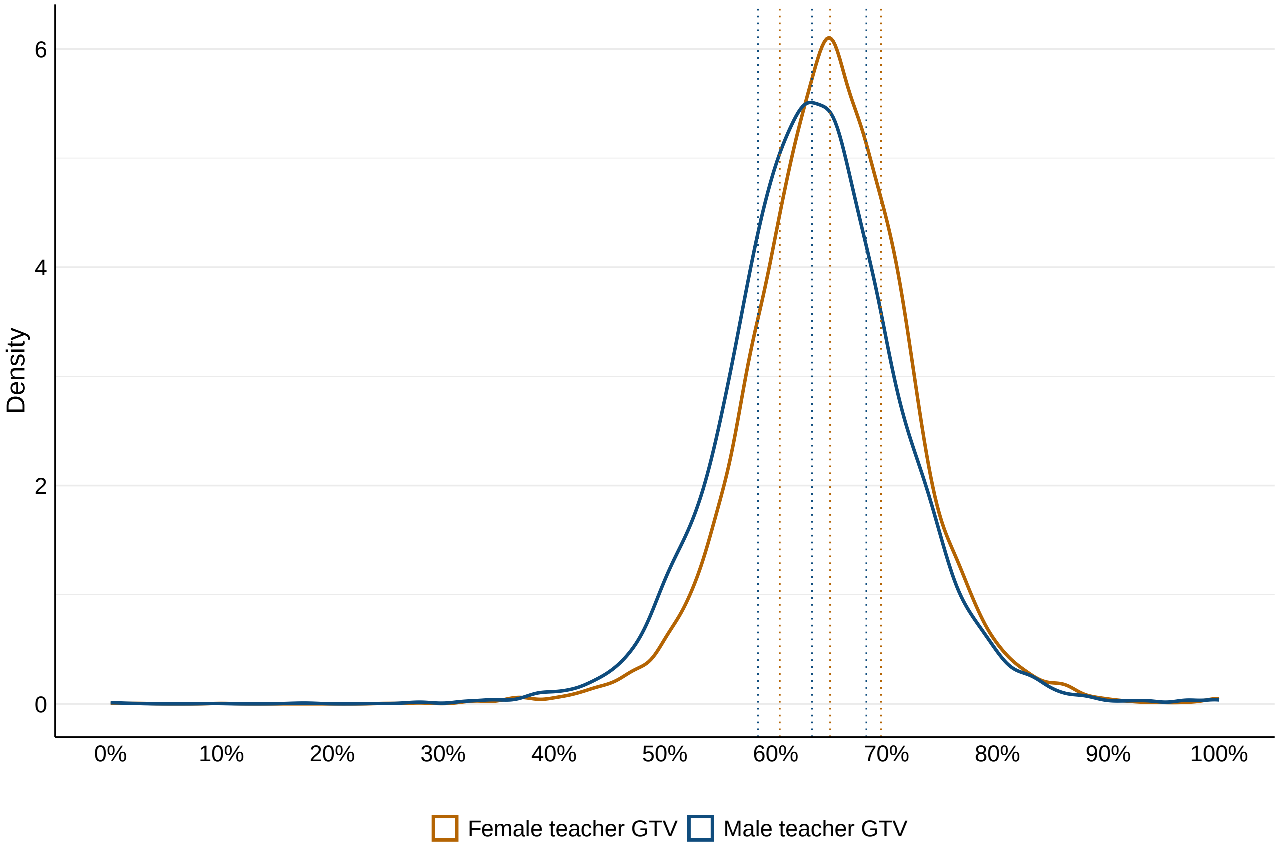
	Positive	**Negative**	**Neutral**
Intellectual	qualité	faille, difficultés, fragilités, incompréhension, étourderie, fragile, erreur	calcul, question, littérale, dst (devoir sur table), contrôle, énoncé, td (travaux dirigés)
Behavioral	sourire, croire, sérieux, studieux, appliquée, exemplaire, accroche, courage, irréprochable, persévérer, bras (ne baisse pas), volonté, fruit (du travail), oser, payer, récompense, sérénité, combativité, discrète, volontaire, épanouie, mérite, autonome, vaincre, tenace, conscience, ténacité, déterminée	confiance (manque de), découragée, panique, stress, pression, effacée, peur, rassure, démotive, réinvestie	dépassée, confiance, hésite, assure, soucieux, poursuivre, subi, continue, encourage, humeur, rectifie
Unclassified	bravo, valide	accident, recul, baisse	éclairci, note, tasse, repère, aide, lien, pose, composition, remet, long, pose, lire, beaucoup, constat, aller, attenue, lève, point, permet, clef, extrême, grand, estompe, repli, relance, multiplier, incite, apparu, cible, sollicite, contribue

C Statistics by Teacher Gender

Table C4 – Share of Male Teachers by Core Subjects

Subject	Share	N	% non-missing
Math	0.58	6,716	0.98
Physics-chemistry	0.58	6,959	0.96
Earth and life science	0.37	5,934	0.84
Philosophy	0.62	6,658	0.97
Modern language 1	0.20	15,176	0.93
Modern language 2	0.18	19,373	0.89

Figure C2 – Distribution of Math Teachers' Leave-one-out GTV – By Teacher Gender



Notes: This figure shows the densities of male and female math teachers' leave-one-out GTV measures. The vertical lines represent the first, second and third quartiles of the GTV distributions.

D Assessing the Randomness of Missing Grade Transcripts

Table D5 – Balancing Test: High School with All Missing Transcripts

	Dep. var: Grade transcripts all missing in high school		
	Coeff.	**S.e**	**p-value**
Female	-0.0375	0.0193	0.1093
Age (years)	0.0898	0.0200	0.0065
Free lunch student	-0.3256	0.0717	0.0062
Foreign student	0.2501	0.2193	0.3056
High SES	0.1077	0.0384	0.0379
Medium-high SES	-0.2715	0.1060	0.0505
Medium-low SES	0.0266	0.0452	0.5817
Rank at DNB: math	-0.2299	0.3815	0.5730
Rank at DNB: math (girls)	-0.0118	0.1368	0.9345
Rank at DNB: math (boys)	0.0829	0.2200	0.7219
Nb. of observations	12,566		

Notes: This table reports the estimation results of a dummy indicating whether the high school is systematically not reporting grade transcripts, regressed on the high school students' average characteristics. Standard errors are clustered at the high school level and are reported in the second column. ***: p-value $<$ 0.01; **: p-value $<$ 0.05, *: p-value $<$ 0.1.

E Teacher Characteristics: Estimation Details and Complementary Results

E.1 Estimating the Teacher Grading Bias

We follow Lavy and Sand (2018) and Terrier (2020) and compute the teacher grading bias as the difference between the class gender gaps in the non-blind (NB) and blind scores (B). We use the (standardized) math grade obtained on the continuous assessment as the non-blind score, and the (standardized) math grade obtained on the *baccalauréat* exam as the blind score. The grading bias (GB) for class c taught by teacher j in year t is therefore defined as follows:

$$GB_{cjt} = \left(NB^{males}_{cjt} - NB^{females}_{cjt}\right) - \left(B^{males}_{cjt} - B^{females}_{cjt}\right)$$

The grading bias assigned to class c is the average bias observed on all classes taught by the same teacher excluding class c i.e., it is the leave-one-out grading bias. A negative (positive) grading bias is indicative of a bias in favor of female (male) students.

The table below reports the average standardized non-blind and blind scores separately for Grade 12 male and female students. On average, female students score above the mean class grade at the continuous assessment, but below when we consider the math *baccalauréat* grade. The reverse holds for male students. The teacher grading bias is calculated as the difference between Columns 3 and 6, and is negative, thus revealing a grading bias favoring female students, both from male and female teachers.

Table E6 – G12 Math Grades at the Continuous Assessment and at Baccalaureat Exam - By Students' and Teachers' Gender

	Boys			**Girls**			
	G12 math (1)	Bac math (2)	Diff. (3)	G12 math (4)	Bac math (5)	Diff. (6)	Teacher bias (7)
All teachers	-0.0133	0.0201	-0.0334	0.0084	-0.0318	0.0402	-0.0736
Female teachers	-0.0277	0.0032	-0.0309	0.0239	-0.0134	0.0373	-0.0682
Male teachers	-0.0025	0.0322	-0.0347	-0.0031	-0.0448	0.0417	-0.0763
Nb. of obs.	21,944	21,923		21,771	21,746		

Notes: This table reports the average standardized math grades obtained at the Grade 12 continuous assessment (Columns 1 and 4) and that obtained at the math *baccalauréat* exam (Columns 2 and 4) separately for male and female students. Columns 3 and 6 report the average difference between both grades. The teacher grading bias reported in the last column of the table reports the average grading bias computed at the teacher level, obtained as the difference between columns 3 and 4. A negative grading bias is indicative of bias in favor of girls.

E.2 Correlation between Teachers' Leave-one-out GTV and their Characteristics

Table E7 – Regression of Leave-one-out GTV on Teachers' Observable Characteristics

	Dep. var: Teacher's leave-one-out GTV (SD)		
	Coeff.	**S.e**	**p-value**
Teacher grading bias one-out (SD)	-0.0365	0.0069	0.0000
Feedback personnalization (SD)	0.0906	0.0071	0.0000
Male teacher	-0.1810	0.0133	0.0000
Head teacher	0.0469	0.0135	0.0005
Nb. of obs.	22,267		

Notes: This table reports the estimation results of the standardized *leave-one-out* GTV measure on other teacher characteristics. Standard errors are clustered at the teacher level and are reported in the second column. ***: p-value < 0.01; **: p-value < 0.05, *: p-value < 0.1.

F Robustness Checks and Additional Results

F.1 Placebo Results: Impact of GTV on other Core *Baccalauréat* subjects

Table F8 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Performance in Other Core Subjects

	All	**Boys**	**Girls**
Grade at baccalauréat (SD): physics	-0.0023	-0.0050+	-0.0001
	(0.0023)	(0.0028)	(0.0029)
Grade at baccalauréat (SD): biology	-0.0022	-0.0037	0.0006
	(0.0025)	(0.0031)	(0.0031)
Grade at baccalauréat (SD): philosophy	-0.0015	-0.0008	-0.0037
	(0.0025)	(0.0029)	(0.0031)
Nb. of observations	602216	317021	282781

Notes: This table reports the coefficients on the standardized *leave-one-out* GTV obtained from the estimation of Equation (6). Each row corresponds to a different linear regression performed on the full analytical sample and separately by gender, with the dependent variable listed on the left. The regression includes high school×elective course×year fixed effects. The number of observations is lower than in the analytical sample because singletons—high school×year×elective cells with the same teacher (and thus identical GTV)—do not contribute to coefficient identification. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.001; **: p-value < 0.01, *: p-value < 0.5, +: p-value < 0.1.

F.2 Robustness Checks

Table F9 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Educational Outcomes-Robustness Checks

	Boys			**Girls**		
	Bsl X	Sh. girls	GTV other	Bsl X	Sh. girls	GTV other
	(1)	(2)	(3)	(4)	(5)	(6)
Academic performance						
Grade at baccalauréat (SD): math	0.0122***	0.0094**	0.0095***	0.0160***	0.0179***	0.0181***
	(0.0026)	(0.0029)	(0.0029)	(0.0027)	(0.0031)	(0.0031)
Type of program ranked first in the ROL						
All Sciences	-0.0036*	-0.0033*	-0.0033*	0.0005	0.0008	0.0007
	(0.0016)	(0.0014)	(0.0014)	(0.0015)	(0.0013)	(0.0013)
Selective Sciences	-0.0019	-0.0033**	-0.0033**	0.0005	0.0011	0.0011
	(0.0015)	(0.0013)	(0.0013)	(0.0011)	(0.0010)	(0.0010)
University Sciences	-0.0016+	-0.0010	-0.0010	0.0010	0.0002	0.0002
	(0.0008)	(0.0007)	(0.0007)	(0.0008)	(0.0007)	(0.0007)
Vocational Sciences	-0.0001	0.0010	0.0011	-0.0014+	-0.0008	-0.0008
	(0.0011)	(0.0009)	(0.0010)	(0.0008)	(0.0007)	(0.0007)
Matriculation in the following year						
All Sciences	-0.0025+	-0.0031*	-0.0030*	-0.0005	-0.0004	-0.0004
	(0.0014)	(0.0012)	(0.0013)	(0.0014)	(0.0012)	(0.0012)
Selective Sciences	-0.0007	-0.0022*	-0.0022*	-0.0004	-0.0002	-0.0002
	(0.0010)	(0.0009)	(0.0009)	(0.0007)	(0.0007)	(0.0007)
University Sciences	-0.0020	-0.0018	-0.0018	-0.0013	-0.0017	-0.0016
	(0.0014)	(0.0012)	(0.0012)	(0.0013)	(0.0011)	(0.0011)
Vocational Sciences	-0.0012	-0.0011	-0.0011	0.0001	0.0006	0.0006
	(0.0011)	(0.0009)	(0.0010)	(0.0007)	(0.0006)	(0.0006)
Nb. of observations	249254	320996	320748	223483	285884	285754

Notes: Each row reports the coefficients on the standardized *leave-one-out* teacher GTV obtained from the estimation of Equation (6) for the different outcomes listed on the first column. It is estimated on the whole sample and separately for Grade 12 male and female students. The regression includes high school×elective course×year fixed effects. The number of observations is lower than in the analytical sample because singletons—high school×year×elective cells with the same teacher (and thus identical GTV)—do not contribute to coefficient identification. Columns 1 and 4 further control for the set of students' baseline characteristics listed in Table 2; columns 2 and 5 control for the average proportion of female students in the classroom, and columns 3 and 6 control for the average *leave-one-out* GTV measured in other subjects for students from the same class. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.001; **: p-value < 0.01, *: p-value < 0.5, +: p-value < 0.1.

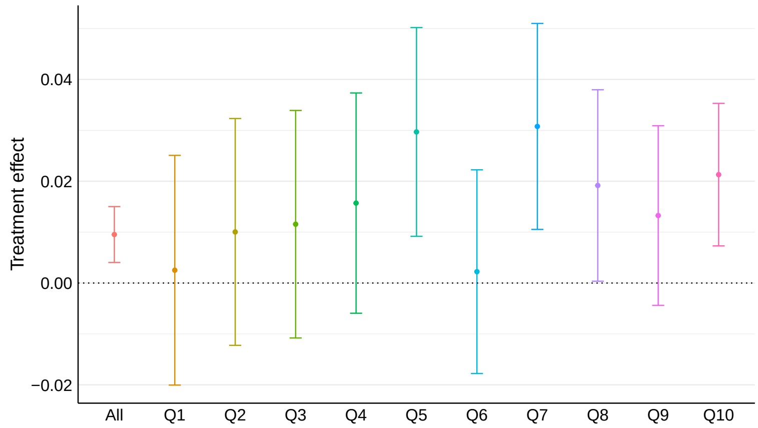
Table F10 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Educational Outcomes - Random Forest

	All	**Boys**	**Girls**
Academic performance			
Grade at baccalauréat (SD): math	0.0148***	0.0100***	0.0196***
	(0.0025)	(0.0029)	(0.0031)
Type of program ranked first in the ROL			
All Sciences	-0.0026**	-0.0037**	-0.0004
	(0.0010)	(0.0014)	(0.0013)
Selective Sciences	-0.0013	-0.0031*	0.0009
	(0.0008)	(0.0013)	(0.0010)
University Sciences	-0.0011*	-0.0015*	-0.0005
	(0.0005)	(0.0007)	(0.0007)
Vocational Sciences	-0.0001	0.0011	-0.0009
	(0.0006)	(0.0009)	(0.0007)
Matriculation in the following year			
All Sciences	-0.0013	-0.0029*	-0.0000
	(0.0009)	(0.0012)	(0.0012)
Selective Sciences	-0.0006	-0.0015+	0.0005
	(0.0006)	(0.0009)	(0.0007)
University Sciences	-0.0009	-0.0012	-0.0010
	(0.0008)	(0.0012)	(0.0011)
Vocational Sciences	-0.0003	-0.0009	0.0004
	(0.0006)	(0.0009)	(0.0006)
Nb. of observations	609331	320996	285884

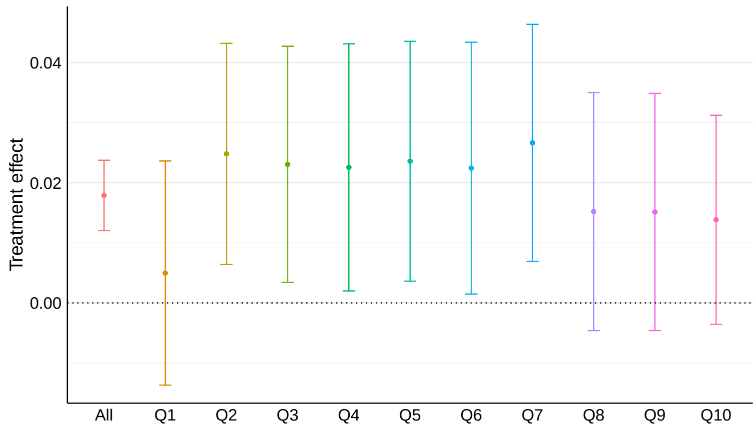
Notes: This table reports the coefficients on the standardized *leave-one-out* GTV obtained from the estimation of a random forest model. Each row corresponds to a different linear regression performed on the full analytical sample and separately by gender, with the dependent variable listed on the left. The regression includes high school×elective course×year fixed effects. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.001; **: p-value < 0.01, *: p-value < 0.5, +: p-value < 0.1.

F.3 Additional Results: Heterogeneity by Initial Math Performance

Figure F3 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Math Performance at *baccalauréat* - By Deciles of Initial Math Performance



(a) Boys

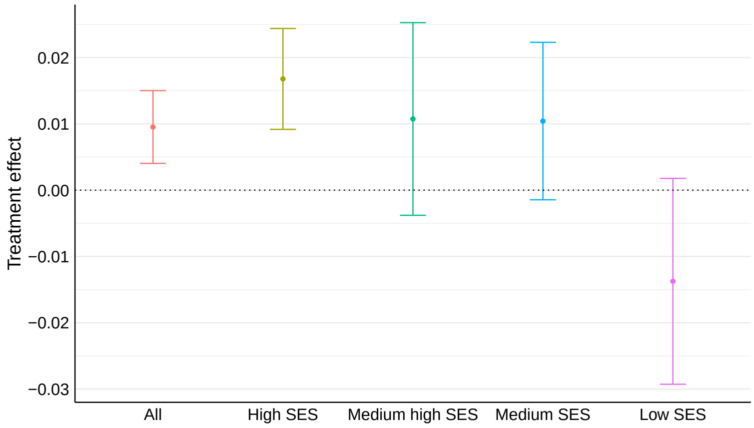


(b) Girls

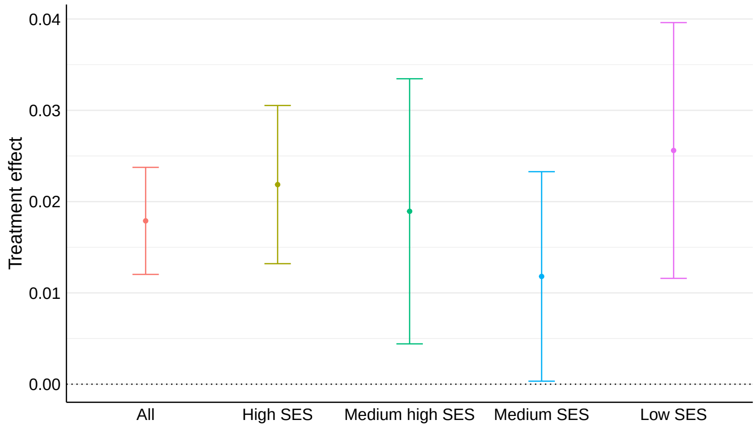
Notes: The figure reports the effect of a one standard deviation increase in *leave-one-out* GTV on students' standardized grade on the math *baccalauréat* exam separately by gender and by initial performance in math. Initial math performance is measured as deciles of percentile rank in math obtained at the DNB nation exam in Grade 9. The solid dots show the estimated coefficients, with 95 percent confidence intervals denoted by vertical capped bars.

F.4 Additional Results: Heterogeneity by Social Background

Figure F4 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Math Performance at *baccalauréat*" - By Social Background



(a) Boys



(b) Girls

Notes: The figure reports the effect of a one standard deviation increase in teacher *leave-one-out* GTV on students' standardised grade on the math *baccalauréat* exam separately by gender and by socioeconomic background. The solid dots show the estimated coefficients, with 95 percent confidence intervals denoted by vertical capped bars.