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

Pauline Charousset, Marion Monnet

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

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

agence nationale de la recherche
AU SERVICE DE LA SCIENCE

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

Pauline Charousset
Marion Monnet

July 2022

Abstract

This paper studies how student gender influences the feedback given by teachers, and how this affects the student's performance in school. Using the written feedback provided to the universe of French high school students by their math teachers over a five-year period, we show that teachers use different words to assess the performance of equally able male and female students. Teachers highlight the positive behavior and encourage the efforts of their female students while, for similarly-performing males, they criticize the students for unruly behavior and praise them for their intellectual skills. To understand how this relates to the student's subsequent educational outcomes, we then match these data to records from French national examinations, as well as these students' higher education application behavior and ultimate institution of enrollment. Using the quasi-random allocation of teachers to classes, we estimate that being assigned to a teacher with feedback that is one standard deviation more gendered improves student math performance by 1.6 percent of a standard deviation on average, but does not affect students' enrollment in higher education 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: French National Institute of Demographic Studies (Ined), Paris, France (e-mail: marion.monnet@ined.fr). We are particularly grateful to Anne Boring, Alex Eble, Julien Grenet and Clémentine Van Effenterre for their numerous reviews and for their wise advice. This paper also greatly benefited from discussions and helpful comments from Elliott Ash, Asma Benhenda, Etienne Dagorn, Marc Gurgand, Elise Huillery, Sylvie Lambert, Arnaud Maurel, Dominique Meurs, Roland Rathelot, Michael Stepner, Camille Terrier, and from all participants at 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. We are especially grateful to the French Ministry of Higher Education for giving us 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 and Chaire Femmes et Sciences is gratefully acknowledged.

1 Introduction

Whether a student's gender influences the feedback received has important implications for gender inequality in education and in the labor market. Students' educational engagement and career choices largely depend on the information about their ability accumulated during their schooling years. This information is inherently noisy as it does not only reflect one's intrinsic ability, but also a range of other factors including assignment difficulty (Landaud et al., 2022), perseverance and effort (Alan et al., 2019), or the way others perceive one's performance (Sarsons, 2019). Differences in how teachers provide feedback to a male or to a female student could distort schooling decisions, and potentially exacerbate gender differences in performance and career choices. For instance, if a student's gender influences the way a teacher delivers feedback, a female student could end up having different beliefs about her ability even if her objective performance is the same as that of a male classmate.

This paper empirically tests whether student gender affects the feedback provided by math teachers in Grade 12 and investigates how it relates to student performance and enrollment choices in higher education. Using the universe of Grade 12 students' transcripts over the period 2012-2017 available in the higher education applications platform, we analyze the written feedback provided by 6,770 math teachers to approximately 700,000 students in France. We find that student gender does influence the feedback provided to equally able male and female students. We further show that students assigned to teachers who use gendered feedback perform better at national examinations but make similar applications and enrollment decisions in higher education. These findings are consistent with a direct effect of performance feedback on motivation, efforts, and therefore achievement, and a negligible impact on longer run outcomes such as self-perception and enrollment decisions.

The specific features of our data allow us to investigate whether student gender influences feedback and relate it to educational outcomes. First, the data contain students' detailed school transcripts thus allowing us to study the written feedback – which is both highly relevant and highly informative to students, and to perform a thorough examination of the wording used by math teachers when assessing male vs. female students' performance. Second, we link students' transcripts to national examinations data to assess how different feedback relates to student performance at high stakes examinations. Third, our data make it possible to test how gendered teacher feedback correlates with other teacher characteristics or teaching practices, including teacher value-added, 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, to compare the educational careers of students exposed to different degrees of gendered feedback.

Leveraging this rich source of information, our first contribution is to study written feedback in 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 correctly predicted observations, 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 female students' performance and male predictors to assess that of males, the better the predictive accuracy, and the stronger the teacher's gender differentiation.

Our first set of results provides evidence that math teachers give differentiated feedback to equally performing male and female students. The average GTV index is 63 percent, meaning that our model correctly predicts gender for 63 percent of students on average. As a point of comparison, this is only marginally lower than the words' predictive power of students' performance level, which is the upper bound of what we could expect given that written feedback's purpose is to assess performance. We document a large variation in the distribution of GTV, suggesting that students are exposed to varying degrees of gendered feedback. To investigate whether the use of a gendered vocabulary is more prevalent in math, we replicate the analysis to five other subjects taken by our pool of Grade 12 students. We further show that math teachers use a more gendered vocabulary than teachers in humanities.

To better understand how the vocabulary used by high GTV teachers differs from gender-neutral ones, 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 different categories reflecting different beliefs and expectations. First, words are classified as either positive, neutral or negative. Second, words referring to students' attitude in class as well as to the effort provision in the subject are classified as "managerial", while those relating to math concepts, the school environment or to students' intellectual ability are classified as "competence-related".[1]

Our second set of results reveals marked gender differences in the nature of the 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 females' behavior and efforts, while two-

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

thirds of male predictors refer to negative managerial aspects. Positive male predictors, however, praise their intellectual skills. Overall, math teachers insist more on positive managerial aspects and encourage the efforts provided by their female students, while equally performing male students are both more criticized for their unruly behavior and 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 investigate how exposure to a high GTV teacher affects students' grade at the national high school graduation exam (*baccalauréat*) in different subjects. Higher education application and enrollment data further allow us to assess the effects of high GTV teachers on students' college application behavior as well as on their actual enrollment outcomes in the year following high school graduation. Our identification strategy exploits the within high school$\times$ elective course variation in GTV and relies on the fact that teacher assignment to classrooms is close to being as good as random conditional on a set of observable characteristics.

Our third set of results shows that being assigned to a teacher with a one standard deviation higher GTV is associated to a 1.6 percent of a standard deviation increase in the performance on the math *baccalauréat* national exam on average, the effect being slightly larger for female students, but not significantly different from that of males (2.1 percent vs. 1.4 percent of a standard deviation). Despite its moderate size, the effect on math performance at *baccalauréat* is larger for students who are exposed to teachers with an above-median GTV. Compared to those exposed to teachers from the lower decile of the teacher-GTV distribution, students exposed to teachers from the fourth decile or above see their math grade increase by up to 6 percent of a standard deviation. The effects on students' top-ranked program in their college application, as well as on their enrollment outcomes are very small in magnitude and mostly not significant.

Finally, we explore the mechanisms that could drive or explain the effect of being exposed to a teacher with higher GTV on math performance. Although we cannot exclude that our measure of gendered feedback is correlated with other teacher characteristics, we try to rule out a set of alternative explanations. First, we show that our results are not driven by other teaching practices such as gender grading bias or feedback personalization. In line with Terrier (2020), we find evidence of a teacher grading bias in math favoring girls but show that it is only weakly correlated with our measure of GTV. Our results are robust to controlling for this grading bias, just like it is robust to controlling for feedback personalization as proxied by a measure of text distance between the teacher's written feedback. We then compute a measure of teacher value-added *à la* Chetty et al. (2014) to investigate whether math teachers with a

high GTV are also better teachers. We find that teacher value-added and GTV are mildly correlated and show that our results are partially channeled through teacher quality. Finally, we investigate whether the type of vocabulary used by teachers has a different effect on male and female students' math performance. We show that the positive effect of exposure to gendered feedback on math performance is reinforced when teachers are more likely to use the vocabulary associated to female students, i.e., teachers who underline the positive behavior and efforts. In line with the feedback and growth mindset literature, we find that females benefit more from the positive managerial feedback than from the negative behavioral feedback and the intellectual praise, while the effects of either type of feedback are not statistically different for males.

This paper contributes to several strands of the literature. It first speaks to the broad literature on performance feedback and individuals' beliefs and choices. This literature has mainly focused on asymmetrical responses to performance feedback, investigating whether individuals adjust their beliefs more after good than after bad news (see for instance the recent contributions of Zimmermann 2020; Coffman et al. 2021). 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, 2021; Bobba and Frisancho, 2020), while other experiments (including lab experiments in psychology) have focused on the nature of feedback and the associated beliefs and expectations conveyed to students. In particular, this literature shows that feedback conveying the idea that intelligence is malleable (growth mindset) enhances students' motivation and attitudes (Corpus and Lepper, 2007), academic performance (Huillery et al., 2021), their sense of belonging as well as their willingness to pursue in the subject (Good et al., 2012), while feedback underlining that intelligence is innate (fixed mindset) has detrimental effects on those outcomes (Canning et al., 2021). While these papers investigate the impact of feedback provided in an experimental setting (either laboratory or field experiment), our paper is the first to document the effect of feedback in a real-life setting.

Our paper also contributes to the literature in social sciences 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, 2020), teaching material (Eble et al., 2021), or the labour market (Ningrum et al., 2020). Our paper uses textual feedback as data to investigate whether the vocabulary used by high school math teachers is gendered. It goes beyond the description of gendered patterns by using the output from the textual analysis, namely the GTV index, in a second step to relate it to students' educational outcomes.

Using non-textual data, another strand of the literature investigates how a person's gender

influences other people's perception of her ability and performance levels. Most of the related papers rely on proxy-matching techniques to investigate the differential treatment of observationally similar individuals and find that women tend to face higher evaluation standards than men from their peers in the labor market or in academia (Sarsons, 2019; Sarsons et al., 2021; Card et al., 2021; Dupas et al., 2021). In this paper, we instead rely on a direct and comparable measure of ability to investigate how gender influences the assessment provided to individuals with similar objective performance levels.

Finally, this paper relates to the literature investigating the scope and impact of gendered teacher behavior on students' 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 at the continuous assessment (Lavy and Sand, 2018; Terrier, 2020), and give them different career advice (Gallen and Wasserman, 2021). These gendered behaviors all have long-term consequences on schooling outcomes. Our paper is, to our knowledge, the first to provide direct evidence of a gendered behavior for another teaching practice, namely written feedback, and to document the short-run effect of being exposed to teachers having different feedback practices on student performance and higher education enrollment decisions.

The remainder of this paper is organized as follows. In Section 2, we provide some institutional background on the French secondary education system and on the admission procedure for higher education. In Section 3, we describe the different data sources that we use, along with some descriptive statistics on the population of Grade 12 students and their math teachers. Section 4 presents our empirical strategy. We detail the different steps to construct our measure of gendered teacher vocabulary (GTV) and to measure the impact of GTV on students' outcomes. In Section 5, we provide a detailed analysis of the gendered vocabulary as well as some statistics on the distribution of our GTV measure. Section 6 shows the impact of being exposed to a higher GTV teacher on academic performance, preferences for higher education programs and enrollment outcomes in the year following high school graduation. Section 7 discusses potential mechanisms and Section 8 concludes.

2 Institutional Background

This section provides some background information on the French secondary education system as well as on the higher education application procedure.

2.1 The French Secondary Education System

In France, secondary education consists of seven years of schooling, divided into four years common to all students and taught in middle schools (*collège*, Grade 6 to 9), and three years of high school (*lycée*, Grade 10 to 12), which provide either vocational or general and technological training. Both the middle and 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 anticipated *baccalauréat* examinations, which include oral and written tests in French, as well as in history and geography for science major students. Students are tested in the remaining subjects at the end of Grade 12. Only students holding the *baccalauréat* can enter higher education.

In general and technological high schools, after a common *Seconde générale et technologique* year (Grade 10), students are tracked into a general (80 percent of students) or a technological curriculum (20 percent of students). General track students further specialize by choosing their major, when entering Grade 11, and their elective course, when entering Grade 12. Students tend to specialize according to both their comparative advantage and their preferences, which leads to marked gender imbalances between majors and elective courses. While female students are slightly underrepresented among science major students (female share: 47 percent in 2018), the economics and humanities majors are largely female-dominated: in 2018, the female share was 60 percent in the economics major and 80 percent in the humanities major (MENJ-MESRI 2019). These gendered patterns in major choice are further reinforced by the choice of elective courses. The differences are particularly striking when focusing on science major students. Female students are largely overrepresented in the earth and life science elective, where they represent 63 percent of students, compared to only 30 percent in computer sciences and 15 percent in engineering. The proportions of female students in math and physics-chemistry electives are more balanced (43 percent and 48 percent respectively).

Gender segregation within French high schools is however limited beyond the segregation induced by the choice of an elective course. The composition of each class is determined by the high school principals who, while taking into account the students' electives when defining the classes, also declare putting gender diversity on top of their priority list (Cnesco, 2015). Most principals also declare valuing some heterogeneity in terms of students' academic achievement level but, unlike for gender, the academic stratification within high schools remains substantial.

2.2 College 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 by providing assistance with the application procedure and some counseling regarding the choice of programs. 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 any program of their choice.

The undergraduate programs students can apply to fall into two broad categories, with, on the one hand, university programs, which are mostly non-selective and open to all high school graduates, and, on the other hand, selective programs. The latter include three different types of curricula, which have a strict academic stratification: 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 prepare students to the entry exam to the most prestigious French colleges (the *grandes écoles*) in science, business, or humanities.

Until 2017, the college admission procedure was centralized through the *Admissions Post-Bac* (APB) online platform for most undergraduate programs.[3] The main round of the procedure implemented a variant of the college-proposing deferred acceptance mechanism (Gale and Shapley, 1962; Roth, 1982). Students were invited 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 list's 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. The grades obtained in the different subjects as well as teachers' written feedback played a crucial role in selective programs' 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 rank in the student's list, but not based on students' grades.

[2]A study by Ly and Riegert (2015) has looked at the determinants of the within high school segregation and finds that the grouping of students according to their elective courses accounts for two-thirds of the observed social and academic segregation.

[3]In 2018, a major reform of the college application procedure allowed universities to select students based on their past academic performance. Since our study is restricted to the period 2012 to 2017, students in our sample were not concerned by this new scheme.

3 Data and Summary Statistics

This section details the different data sources used to build our measure of gendered teacher vocabulary (GTV) and to quantify the effect of being exposed to a teacher with higher GTV on students' outcomes (Section 3.1). We also present summary statistics on the sample of Grade 12 science major students and on their math teachers (Section 3.2).

3.1 Data Sources

We use three main administrative databases: the college application data for six cohorts of Grade 12 students (2012-2017) 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 of information is the comprehensive application data from the APB platform over the period 2012–2017. A substantial amount of information is collected by this platform during the application process. First, we use the students' digitalized academic records to retrieve teachers' written feedback on all the subjects taken by Grade 12 students (two trimesters). This is the main input used to build our measure of gendered teacher vocabulary. These transcripts also contain the students' grades at the continuous assessment for both Grade 11 and 12. Teachers and students are uniquely identified in the data, which enables us to link the transcripts to the characteristics of students and teachers that are contained in a separate APB file. Along with basic sociodemographic information (gender, place and date of birth, parental socio-economic status, etc.), the APB data provide detailed information on students' high school outcomes (school track, major and elective choices), as well as information on the teachers' gender, subject taught, and head teacher status.

The APB data further keep a record of the final rank-order list of programs submitted by each applicant, the matching outcome, i.e., the program to which each student was admitted, along with the students' acceptance decision (acceptance, conditional acceptance or 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 the grades obtained by students in two national examinations: the *Diplôme National du Brevet* (DNB), taken at the end of Grade 9, and the *baccalauréat*, taken at the end of Grade 12, both of which are anonymously and externally graded. We use the former

to control for the students' past academic performance in the estimation procedure, while the latter is used as our main measure of student performance at the end of high school. To make grades comparable across years, we transform the initial grades ranging between 0 and 20 into percentile ranks, where 0 and 100 are respectively the ranks for the lowest and the highest performing students.

College enrollment data. To track Grade 12 students' enrollment outcomes in the following academic year, we use the *Système d'Information sur le Suivi de l'Étudiant* (SISE), which is managed by the Statistical Office of the French Ministry of Higher Education. This dataset, which covers the academic years 2012 to 2017, records all students enrolled in the French higher education system outside of CPGE and STS, except for the small fraction of students enrolled in undergraduate programs leading to paramedical and social care qualifications. For selective programs, we use a separate administrative data source called *Bases Post-Bac*.

Sample restrictions. Given that the focus of this study is on math teachers' feedback, we restrict our sample to students enrolled in the science major in Grade 12, as they are the ones interacting most frequently with their math teachers, compared to students in the humanities or the social sciences majors.[4] These students are also the most likely to opt for a science major in college and may therefore be more responsive to their math teacher's feedback. We exclude students for whom the math teacher's identifier or the grade transcript is missing, which represents about 50 percent of Grade 12 students from the science track in 2012, down to 15 percent in 2017 (Table 1). In the vast majority of cases (between 70 and 95 percent of missing observations), teachers' identifiers and grade transcripts are missing because the entire high school is not reporting its students' grade automatically on the APB platform. Dropping these observations therefore amounts to dropping entire high schools hence does not constitute a threat for the internal validity of our analysis.[5] Finally, we restrict our sample to high schools having at least two science major classes, since our identification strategy relies on a within-school comparison of students (see Section 4), and to teachers who have taught at least two classes over the period 2012–2017. These restrictions remove between 6 and 20 percent of students. Depending on the year considered, the sample of analysis includes 40 to 75 percent of Grade 12

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

[5]It might, however, affect the external validity of our analysis. Table D5 in Appendix D shows the OLS coefficients of a dummy indicating whether the high school has all grade transcripts missing regressed on the high school's average characteristics. High schools with a higher share of 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 only marginally affects the probability of not reporting grade transcripts.

science major students, for a total of approximately 700,000 observations.

3.2 Summary Statistics

Students. Table 2 provides summary statistics of Grade 12 science major students' characteristics for the whole sample of analysis, separately for male and female students. Students are 18 years old on average and mostly come from a high (43 percent) or a medium-high (16 percent) socio-economic background.[6] Female students are slightly underrepresented in the science major as they account for 47 percent of science major students but 54 percent of all general track Grade 12 students (MENJS-MESRI, 2018). Turning to elective courses, we note striking gender differences. Half of female students opt for the earth and life science elective compared to only one fourth of males. 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 between male and female students relates to their past academic performance, as measured by the national percentile rank on the DNB exam. Males' average rank in math is approximately four points above that of females. On the other hand, females outperform males in French at both the DNB and *baccalauréat* exams, with an average percentile rank that is 10 points higher than that of their male peers. Both imbalances, in terms of elective choices and past school performance, are accounted for in our identification strategy and estimation procedure.

Math teachers. Table 3 reports some descriptive statistics for the sample of math teachers in Grade 12 Science major. There are 6,772 math teachers in the sample, of whom 58 percent are males. A little more than half of these math teachers have been the head teacher of a class at least once during the period covered by our data. Those teachers are likely to have a stronger influence on students' performance and enrollment behavior as they counsel students on top of teaching them. Each teacher is in charge of only one Grade 12 science major class on average each year, with an average class-size of 28 students (90 percent of teachers teach only one Grade 12 science major class per year). Teachers appear almost four times in the sample, meaning that we have on average four classroom observations per teacher, which is crucial for the reliability of our GTV measure (see Section 4). Finally, the average length of a teacher's feedback for a given student is made of 12.5 words, with large variability across teachers.

[6]Students' socioeconomic status (SES) is measured using the French Ministry of Education's official classification, which uses the occupation of the child's legal guardian to define four groups of SES: high (company managers, executives, liberal professions, engineers, intellectual occupations, arts professions), medium-high (technicians and associate professionals), medium-low (farmers, craft and trades workers, service and sales workers), and low (manual workers and persons without employment).

4 Empirical Strategy

The first part of this section describes the estimation procedure that we implement to measure gendered teacher vocabulary (GTV) (Section 4.1). The second part presents the identification strategy to estimate the effect of being exposed to a teacher with higher GTV on students' educational outcomes (Section 4.2).

4.1 Measuring Gendered Teacher Vocabulary (GTV)

The measure of teacher GTV proposed in this paper leverages the rich data on teachers' written feedback provided to students in their Grade 12 academic records. This feedback reflects the teachers' perception of students' performance, work, and behavior in class throughout the year. It is both highly relevant and highly informative to students: it is provided three times a year to students, is shared with their parents, and is considered by selective higher education programs during the college application process. Therefore, the way feedback is framed may considerably influence students' behavior and outcomes. To investigate whether the words used to characterize a students' work, behavior and ability differ by gender, we build a model that predicts students' gender based on the words used in the teachers' feedback. 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' prior academic performance. We then use this fitted model to compute a measure of gendered vocabulary for each teacher based only on the classrooms that he or she has taught. The different estimation steps, which are inspired from the text mining literature (Gentzkow et al., 2019), are presented below while the detailed procedure can be found in Appendix A.

Data preparation. The first step consists in converting the corpus of teacher feedback into a statistical database, which is performed in two steps. First, we rely on text mining techniques to replace each word by its root and hence ensure its gender neutrality. Second, the corpus of teachers' feedback is converted into a matrix that contains one row per feedback and W_n columns, where W_n denotes the number of distinct words appearing in the corpus. Each of these columns is a dummy that takes the value one if the considered word appears in the student's feedback, and zero 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 to 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:[7]

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

In practice, the model's predictive quality, as measured by the accuracy – the proportion of correctly classified observations – could be influenced by two factors that we seek to neutralize before any estimation or prediction is done. First, since gender is correlated with math performance (see Section 3.1), Model (1) is likely to perform better on classes with stronger gender imbalances in math. To alleviate this concern, we undersample for each teacher as many boys and girls from each quartile of prior math ability (as proxied by the DNB percentile rank in math). This ensures balanced training and hold-out samples, which are made of 50 percent of male and female students from each ability level. Second, feedback length varies substantially among teachers (see Table 3), implying that teachers writing lengthier feedback have mechanically higher accuracies. For feedback with an above-median length, we circumvent this issue by randomly sampling 12 words, i.e., the median number of words.

Estimating Model (1) on the balanced subsample yields a student gender accuracy of 63 percent. In other words, our model correctly predicts students' gender in 63 percent of cases. It performs slightly better at predicting male students' gender than female students' (65 vs. 63 percent).[8]

[7] A student is classified as female if her predicted probability of being a female is larger than 0.5. Otherwise, she is classified as male.

[8] We tried more flexible specifications of the model by adding interactions between words (*bigrams*) as predictors, in addition to single words (*unigrams*). This more complex specification did not improve the predictive quality of the model.

Gendered teacher vocabulary (GTV). We define each teacher j's GTV for class c as the share of her students whose gender is correctly predicted by the model, i.e., GTV is the predictive accuracy of the model fitted on her sample of students. As teacher j's estimated GTV for class c could capture some unobserved class-specific gender differences in behavior or performance, we compute an alternative measure that we call the *leave-one-out* GTV, which is defined as the average of teacher j's GTV over all the classes she taught during the study 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:

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

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

Both GTV measures theoretically lie between 0 (the model systematically misclassifies females as males and males as females) and 100 (all students are assigned their actual gender). The higher the accuracy for a given teacher, the better we can, on average, recover his or her students' gender based on the words she uses in her feedback, and hence the stronger the gender differentiation in the vocabulary used in her assessments. A model that randomly assigns each student a gender with probability 0.5 would achieve a 50 percent accuracy, meaning that our model predicts gender better than random guessing for all teachers whose accuracy is above 50 percent[9].

4.2 Identification Strategy

Besides documenting gendered practices in teachers' written feedback, the second objective of this paper is to characterize the relationship between the GTV measure presented above and students' performance and enrollment outcomes. Our identification strategy relies on the comparison of students enrolled in the same high school, in the same elective course, but who are

[9] An accuracy below 50 percent is possible in our setting given that the prediction is performed on small samples at the teacher level. However, accuracies are averaged over 100 estimations to limit such random fluctuations, and the *leave-one-out* GTV is itself an average of multiple accuracies, therefore reducing the noise inherent to the measure.

exposed to math teachers with 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 as we use the *leave-one-out* GTV described in Equation (5). The coefficient γ_{set} is a set of high 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 being assigned a teacher with a one standard deviation higher GTV. The standard errors are robust and clustered at the teacher level.[10] The validity of our identification strategy requires that the *leave-one-out* GTV is not systematically correlated with students' characteristics. We formally test this in Section 6.

5 Math Gendered Feedback

In this section, we first describe the distribution of our GTV measures (Section 5.1) and present a qualitative analysis of the gendered vocabulary used by teachers (Section 5.2).

5.1 Distribution of the GTV measures

Figure 1 shows the density and the cumulative distributions of the GTV and *leave-one-out* GTV measures separately. It provides evidence of a correlation between students' gender and the feedback received in math, controlling for students' prior math ability. Our model predicts gender better than random guessing for 90 percent of math teachers if we consider the GTV measure and for over 95 percent with the *leave-one-out* GTV.[11] For the median teacher, 63 percent of students' gender is correctly predicted. As a point of comparison, the model achieves a median accuracy of 66 percent when predicting the students' math performance, which is the upper bound of what we could expect given that the feedback aims at assessing performance.[12] When breaking down the GTV distributions by teachers' gender, we note that,

[10] Although a preferable approach would be to bootstrap standard errors to account for prediction error, we do not implement this correction due to computational limitations.

[11] In total, only 4 percent of teachers have a leave-one-out GTV below 50 percent, the vast majority of which falls between 44 percent and 50 percent. This is mostly explained by the fact that the corresponding teachers are observed three times on average, compared to 4 times for other teachers. Their leave-one-out GTV is therefore slightly noisier.

[12] When predicting student's performance, the response variable is equal to one if the student is among the top 50 percent performers of his class at the math DNB exam and zero otherwise.

on average, female math teachers differentiate their vocabulary slightly more than their male colleagues (see Figure C3).

To assess whether using a gendered vocabulary is specific to math, we replicate the procedure for the following core Grade 12 science track subjects: physics & chemistry, biology, philosophy and modern language 1 and 2. Figure 2 displays the leave-one-out GTV distributions for the different subjects. It shows that the *leave-one-out* GTV distribution for humanities-related subjects is shifted to the left compared to science-related subjects.[13] This suggests that teachers in philosophy and modern languages are, on average, less likely to use a gender-specific vocabulary in their feedback compared to math, physics and chemistry teachers, while biology teachers are somewhere in-between.[14] Philosophy is a particularly relevant point of comparison, since the gender composition of Philosophy teachers is close to that of math-intensive subjects (62 percent of male teachers, see Appendix Table C4). Yet, philosophy teachers seem to use a more gender-neutral vocabulary.

5.2 Qualitative Analysis of the Best Gender Predictors

Definition of the classification. A high degree of gender differentiation in the vocabulary used, as measured by a high GTV, can convey different teachers' beliefs and expectations. It may reflect gender stereotypes regarding students' math ability, but it could also be that the teacher adapts her feedback to the different student profiles she perceives. For example, as shown by the growth mindset literature, female students benefit more from feedback insisting on their effort rather than on their ability, which could be a reason for the teacher to differentiate her vocabulary (Corpus and Lepper, 2007; Good et al., 2012; Canning et al., 2021).

To better understand the extent to which the gendered vocabulary translates into gendered beliefs and expectations, we explore the actual feedback content by analyzing the best student gender predictors. Building on the psychology literature on the classification of teacher feedback and mindsets (Morgan, 2001; Burnett, 2002; Dweck, 2006), we classify the student gender predictors into five different categories to capture the different beliefs and expectations conveyed. First, depending on the valence of the word, we assign it a positive, neutral or negative category. Second, words referring to students' attitude in class as well as the efforts provided for the subject are classified as "managerial", 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

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

[14]The median leave-one-out GTV are as follows: 63.4 percent in physics and chemistry; 62.7 percent in biology; 60.1 percent in philosophy; 59.3 percent in modern language 1 and 59.8 percent in modern language 2.

either of the two categories remain unclassified.[15] The classification of the top 100 male and female predictors is displayed in Appendix Tables B2 and B3.[16]

Analysis of best gender predictors. The analysis reveals marked differences in the qualifiers used by teachers depending on the gender of the student. Figure 3 reports the odds ratios derived from the estimation of the model described by Equation (1), for the top 10 predictors of each gender. A feedback mentioning the student's lack of confidence, her propensity of getting discouraged or her cheerful aspect ("smiling") is between 1.8 and 2.1 times more likely to be directed to a female than to a male student, relative to a feedback that does not mention it. Teachers are also more likely to mention that female students are stressed or panicked, and to insist on their exemplary conduct ("exemplary", "studious"). On the other hand, a feedback describing the student as childish ("childish", "has fun"), insisting on the need for careful handwriting, or praising the student's curiosity and intuitions is between 1.8 and 2.3 times more likely to be received by a male rather than by a female student, relative to a feedback that does not mention these terms.

Figure 4 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, managerial or unclassified words. The first striking feature is the relative proportions of positive versus negative type of feedback by gender. Among the top 30 male predictors, only 8 correspond to a positive feedback, while roughly two thirds of the best female predictors can be considered as positive. Most interestingly, conditional on being positive, the best male predictors almost all qualify the student's competence-related aspects ("curious", " idea", "interest", "intuition"), while nearly all of the best female predictors qualify managerial aspects ("irreprochable", "willingness", "persistent"). On the other hand, more than 75 percent of the best male predictors can be classified as negative and the vast majority refer to a disruptive behavior ("has fun", "childish") or to a neglected work-effort ("waste", "superficial").[17]

[15] We attempted to confirm our classification with data-driven techniques using bi-term topic models tailored for short texts, but these models performed poorly on our data. Our data are indeed quite specific in that texts are very short, with an average of 12 tokens per teacher, the overall vocabulary is quite limited (on average 1,600 words), with little variation in the topics used (they almost all relate to academic performance and behavior). We therefore faced the typical challenges inherent to such short texts: the generated topics 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] Even though every token has been classified, we only show the top 100 predictors given that other predictors are not more frequently used for female or male students (their odds ratio is around 1) and cannot be classified in any of the five mentioned categories in the vast majority of cases.

[17] The classification of the top 30 gender predictors when bigrams are used instead of unigrams is displayed in Appendix Figure B1, and leads to the same conclusions.

The above results are confirmed when we consider all gender predictors. Figure 5 shows the proportions of negative, positive and neutral feedback, conditional on feedback being related to competence (Panel A) or behavior (Panel B). Panel A shows that among female predictors that can be classified as competence-related, only 20 percent correspond to a positive feedback compared to 38 percent for male predictors, while 17 percent (11 percent for male students) are negative, the rest being neutral. Symmetrically, among the female predictors that can be classified as managerial, 44 percent correspond to a positive feedback compared to 29 percent for males. The latter receive a much larger share of negative feedback: as much as 43 percent of managerial male predictors are negative while this proportion is only 31 percent for females.

Turning to the proportions of competence, managerial and neutral words conditional on having a positive or a negative feedback (Panel B), we observe that conditional on being positive, top female predictors qualify their competence-related skills in only 17 percent of cases compared to 39 percent for their male counterparts. Regarding the breakdown of negative predictors, 38 percent of negative male predictors relate to managerial matters compared to 36 percent for female predictors, and those proportions are respectively 16 percent and 9 percent for negative competence-related predictors.

Vocabulary used by teachers' decile of GTV. We further investigate whether teachers with varying degrees of GTV differ from each other, by comparing teachers' gender gaps in the share of positive words among competence and managerial-related feedback by decile of GTV (see Figure 6). Panel (a) plots the absolute values of the teachers' gender gaps while Panel (b) displays the share of teachers having a gender gap in favor of females, separately for competence and managerial-related feedback. The gender gaps in the share of positive words increase at a growing pace with GTV deciles, indicating that teachers with a higher GTV tend to provide relatively more positive feedback to one gender over the other, and even more so as the GTV decile is high. This is true for both managerial-related and competence-related feedback, with a gender gap that goes from 6 to 7 percentage points in the lower GTV deciles up to 10 points in the 10^{th} decile. In line with the findings from the analysis of the gender predictors, Panel (b) reveals that higher GTV teachers are overwhelmingly more positive towards female students in their managerial-related feedback, and more negative in their competence-related feedback.

Taken together, these descriptive statistics indicate that teachers do use a differentiated vocabulary for their male and female students. They seem to insist more on positive managerial aspects and to encourage efforts for their female students, while equally performing males are

both more likely to be criticized for their unruly behavior and to be praised for their intellectual skills. In the following section, we investigate how this gendered feedback affects students' performance, future choices and enrollment outcomes.

6 Impact of Gendered Feedback on Students' Outcomes

Having documented differences in Grade 12 math teachers' gendered vocabulary, we turn to the impact of GTV on students' outcomes. First, we perform a series of statistical tests aimed at validating our empirical strategy (Section 6.1). We then discuss how the exposure to teachers with different levels of GTV affects academic performance, college application behavior and enrollment the year following high school graduation (Section 6.2). We show that our results are robust to a series of alternative specifications.

6.1 Validity of the Empirical Strategy

6.1.1 Exogeneity Assumption

The validity of our identification strategy requires that teacher GTV is not systematically correlated with students' characteristics. Ideally, we would want teachers to be randomly allocated to classes within a high school for a given elective course. We formally test this below.

Balancing tests. Table 4 reports the coefficients from a regression of the teachers' standardized *leave-one-out* GTV, defined at the class level, on students' socio-economic characteristics and baseline academic performance, along with a set of high school×elective course×year fixed effects. The table shows that teacher GTV is not systematically correlated with students' observable characteristics. Out of the twelve characteristics included in the regression, only the "foreign student" dummy and the percentile rank on the written French *baccalauréat* examination are marginally significant, and the magnitude of these coefficients is very small.[18] This test provides evidence in favor of teachers' random allocation conditional on high school×elective course×year fixed-effects.

Random allocation of students. To check whether students are randomly allocated to teachers within a given high school, elective course and for a given year, we perform a series of

[18]The coefficients can be interpreted as follows: a 1 percentage point increase in the share of foreign students is associated with a 1.26 percent of a standard deviation increase in teacher GTV, while a 10 percentile rank increase in the average rank on the French baccalauréat is associated with a 0.1 percent of a standard deviation increase in teacher GTV.

Pearson's Chi-square tests of independence. For each unique combination of high school, elective course, and year, we tabulate math teachers' identifiers with each of the students' baseline characteristics and test for independence.[19] Table 5 reports the percentage of p-values below the nominal values of 0.05 and 0.01. Except for the female dummy, we find that the empirical p-values are close to the nominal values (between 4.5 percent and 8 percent of p-values are below nominal levels). For the female dummy, the empirical p-value is 11 percent, suggesting that in 11 percent of 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 presented in Section 6.2 are not driven by the 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.

Overall, the tests performed in this section suggest that in a given high school, elective course and for a given year, students are close to being randomly allocated to classes.

6.1.2 Reverse Causality

A legitimate concern regarding the GTV measure is that teachers' behavior could be influenced by the type of students they are exposed to. In this case, our measure would not pick up some stable trait in the teachers' gendered vocabulary. Below, we argue that this type of reverse causality is unlikely to be an issue in our setting.

First, we have shown in Table 4 that students' observable characteristics are rather well balanced across the distribution of teacher GTV: teachers who are more or less differentiating their vocabulary are not systematically assigned a specific type of students.

Second, each class is assigned its teacher's *leave-one-out* GTV measure, i.e., the average GTV measured in all the classes ever taught by the teacher except the considered class, which ensures that students do not contribute to the GTV measure they are being assigned.

Third, looking at the distributions of *leave-one-out* GTV measures estimated for other subjects further highlights the specific nature of science-related subjects, for which students' gender is better predicted on average than for humanities-related subjects (see Figure 2). Students' gender is correctly predicted in 59 percent of cases for humanities-related subjects, against 63 percent for math or physics and chemistry on average. This suggests that, on average, for a given class, science teachers differentiate more their vocabulary by gender. Our measure

[19] Continuous baseline characteristics such as age are previously dichotomized. The resulting variables take the value 1 if the student is above the median value and 0 otherwise. Measures of academic performance such as the students' percentile rank on the DNB examination are transformed into quartiles.

therefore captures differences that go beyond class-specific characteristics.

Finally, the fact that teachers' GTV is computed for multiple years and classes offers the opportunity to measure whether teachers' GTV is persistent across classes and over time. The correlation between a teacher's GTV and *leave-one-out* GTV, i.e., the correlation between a given GTV and its average computed in other years×classes, is 0.161 and is statistically significantly different from zero.[20] It is worth noting that as we are correlating several GTV measured with error because of the small sample size used to form the prediction at the class level, this correlation suffers from an attenuation bias. As a comparison, in the teacher value-added literature, the within-teacher correlation is usually around 0.3 (Chetty et al., 2014). Overall, we are confident that our GTV measure captures persistent differences between teachers' GTV.

6.2 Effect of Exposure to Gendered Teacher Vocabulary on Performance and Enrollment

Academic Performance. Panel A of Table 6 reports the estimated effect of being exposed to a teacher with a higher *leave-one-out* GTV on students' standardized math performance on the *baccalauréat*, based on Equation (6). The results indicate that a being exposed to a teacher with a one standard deviation higher GTV raises math performance at *baccalauréat* by 1.6 percent of a standard deviation on average, significant at the 1 percent level. This effect corresponds to moving from a teacher with an average GTV to a teacher from the 86^{th} percentile of the GTV distribution. The effect is slightly larger for female students, whose math grade increases by 2.1 percent of a standard deviation when exposed to a teacher with a one standard deviation higher GTV, than for male students (1.4 percent of a standard deviation). The effects, however, are not statistically different by student gender. As placebo tests, Appendix Table E6 reports the effect of math teacher GTV on the standardized grades obtained in physics, biology and philosophy on the *baccaularéat* exam. The effects are not statistically significant for each of these core subjects, which is consistent with our baseline estimates capturing the effect of the math teacher rather than by unobserved differences between classes.

These moderate effects hide heterogeneous responses depending on the degree of gender differentiation in the math teacher's feedback. Instead of including the teacher GTV linearly in the equation, we explore the intensity of the treatment by regressing students' math grade on a set of GTV deciles. The first (last) decile corresponds to the bottom (top) 10 percent of the math teachers' leave-one-out GTV distribution. Figure 7 plots the coefficients associated

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

with the GTV deciles along with their 95 percent confidence intervals, separately for male and female students. Compared to students exposed to the bottom 10 percent of teachers in terms of GTV, those exposed to teachers from the 4^{th} decile or above see their *baccalauréat* performance increase by a significant 4 to 6 percent of a standard deviation on average, for both males and females. By contrast, we find no evidence of significant heterogeneity by students' prior math performance or socio-economic status (see Appendix E for details).

Results on college applications and enrollment. Although being exposed to a teacher with a one standard deviation higher GTV is found to significantly improve female and male students' performance in math, we find no evidence of significant effects on their college application and enrollment outcomes. Treatment effects on the probability that students rank a STEM program as top choice in their college applications (Table 6, Panel B) and of enrollment in a STEM undergraduate program in the year following high school graduation (Table 6, Panel C) are small and statistically insignificant at conventional levels. If anything, male students are only marginally less likely to top rank and enroll in a selective STEM program (-0.4 percentage point), which represents a 1.9 percent decrease from the baseline of 21 percent. We find no evidence of any heterogeneous effects by deciles of teacher GTV, by student's prior math performance or by student's socio-economic status.

Robustness checks. Our results are robust to a series of sensitivity tests reported in Appendix Table E7. First, to check that our results are not driven by the slight imbalances in the share of foreign students and in the rank at the written French exam mentioned in Section 6.1, we estimate the model described by Equation (6) controlling for students' baseline characteristics (columns 1 and 4) and for the share of female in the classroom (columns 2 and 5). Second, to further make the case that the effect we estimate is specific to the math teacher GTV, we control for the average GTV measured in the same classroom for the five other core subjects (Columns 3 and 6). Including any of these controls does not change the magnitude nor the significance of the results. The effects of a one standard deviation increase in GTV on math performance range between 1.2 and 1.4 percent of a standard deviation for boys, and between 1.8 and 2.1 percent for girls. The limited, yet significant, effects on the probability to top-rank a STEM program in the ROL, as well as on the probability to matriculate in a STEM program are still there for boys and represent approximately a 0.5 percentage point decrease.

7 Mechanisms

As our setting does not involve feedback manipulation, but only manipulation in exposure to teachers with different levels of gendered feedback, our GTV measure is potentially correlated with other teacher features. The estimated coefficients could therefore capture the effects of these related characteristics. In this section, we explore the mechanisms that could potentially drive the effects of math teachers' GTV on math performance and show that the effects of exposure to gendered feedback remain after accounting for those potential confounders. First, we investigate whether teachers using a gendered vocabulary are also encouraging females by overgrading them relative to males (Section 7.1). Second, we investigate whether teachers using gendered feedback are also more likely to personalize their students' feedback (Section 7.2). Third, we compute a measure of teacher quality to assess whether math teachers with a higher GTV are also better teachers (section 7.3). Finally, we test whether the effect on math performance are different when students are exposed to teachers more likely to use the male- or female-specific vocabulary for a given level of gendered feedback (Section 7.4).[21]

7.1 Teacher Grading Bias

A first mechanism through which teachers could encourage girls and increase their performance is by overgrading them relative to their male peers. This teacher grading bias in favor of female students and its positive impact on female students' school performance and enrollment choices have already been documented, e.g., Lavy and Sand (2018) and Terrier (2020). Using a similar approach, we estimate teachers' grading bias by taking the difference between the gender gap in math test scores at the continuous assessment and the gender gap on the math *baccalauréat* exam (see Appendix F for details). A negative (positive) grading bias is indicative of a grading bias in favor of females (males) on the continuous assessment. Consistent with the literature, we find that, on average, high school math teachers exhibit a grading bias in favor of female students (Table F8).

Turning to the correlation between the grading bias and our measure of GTV, Panel (a) of Figure F8 shows that a one standard deviation increase in GTV is associated with a -0.06 standard deviation decrease in the grading bias, significant at the one percent level. This correlation is low but suggests that teachers who use a more gendered vocabulary are also slightly more likely to encourage females through higher continuous assessment grades. However,

[21] Given the low and mostly non-significant effects of GTV on higher education choices and enrollment outcomes, we only show the analysis of the mechanisms for the standardized grades in mah. Results for other outcomes are available upon request.

controlling for the grading bias in the main specification does not affect the magnitude of the GTV effect. Panel a of Table 7 reports the coefficients estimated on GTV and on the teachers' grading bias. It first shows that a one standard deviation increase in teacher grading bias increases math performance by 1.4 percent of a standard deviation for boys and 1.2 percent for girls, a magnitude that is comparable to our effects of exposure to gendered feedback. However, its inclusion as a control does not change the estimated coefficients on GTV. We can therefore exclude teachers' grading bias as a first-order mediator of the impact of GTV.

7.2 Teacher Feedback Personalization

Next, we investigate whether gendered feedback influences students' outcomes beyond feedback personalization, which has been shown to improve students' motivation and to raise their performance (Koenka and Anderman, 2019). Feedback personalization is likely to be related to GTV, as a lower degree of feedback personalization may mechanically decrease the chances that the model detects gendered feedback. To take an extreme example, a teacher who copy-pastes his feedback for students falling in the same grade range leaves no chance to identify gendered patterns. If this is the case, then GTV and feedback personalization are likely to be positively correlated and the GTV coefficients on performance might partly capture the effects of feedback personalization.

To investigate that matter, we build a proxy for teacher feedback personalization and compute a measure of within-teacher text distance. For each teacher, we compute the euclidean text distance between the feedback provided to each of his students.[22] The larger the text distance between the different feedback, the more this teacher personalizes feedback. Figure F6 displays the distribution of teacher text distance as well as its correlation with the leave-one-out GTV. While significant, the figure clearly shows that the relationship between our proxy for feedback personalization and the use of a gendered vocabulary is very weak.

We find that controlling for the within-teacher text distance in the estimation of the model described by Equation (6) does not affect the estimated relationship between GTV and student performance. Panel b of Table 7 shows that while being exposed to a teacher personalizing his feedback seems beneficial to both boys and girls, exposure to a teacher using gendered feedback still significantly improves math performance by 1.4 percent of a standard deviation for boys and 2.1 percent for girls. Therefore, we can rule out the idea that being exposed to a teacher

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

with higher GTV affects performance through feedback personalization rather than through gender differentiation.

7.3 Teacher Quality

We then investigate whether the positive effects resulting from exposure to a higher GTV could be mediated by differences in teacher quality. We compute a measure of teacher value-added following the methodology described in Chetty et al. (2014). We start by regressing the standardized math *baccalauréat* grade on a set of students' baseline characteristics, measures of prior academic performance, and teacher fixed effects. We predict residuals and use them to compute the average residualized test scores for each class×year combination. Class residuals in year t are regressed on their lags and leads, whose coefficients are the shrinkage factors. Finally, the coefficients obtained are used to predict teachers' value-added in year t. All the details of the estimation as well as the distribution of teacher value-added can be found in Appendix F.

We then check whether teachers' GTV is correlated with their quality. Panel (b) of Appendix Figure F8 suggests a small yet significant quadratic relationship, where teachers with GTV measures that are two standard deviations below or above the average have a slightly lower value-added.

Controlling for teacher quality in Equation (6), we find that the effect of exposure to gendered feedback on math performance is reduced by 0.6 percentage point for boys and by 0.7 percentage point for girls compared to the specification without this control. It goes from +1.4 percent of a standard deviation in math grade to +0.8 percent for males, and from 2.1 percent to 1.4 percent of a standard deviation for females (Table 7, Panel c). Although it is reduced, the effect of exposure to higher GTV on students' math performance remains significant, which suggests that our results are only partially channeled through teacher quality.

7.4 The Effect of Male-Specific or Female-Specific Vocabulary

The last mechanism we explore is whether the type of vocabulary used by teachers, i.e., the words more likely to be used for female or for male students, triggers different responses from students. While it is possible to classify and describe the general female-specific and male-specific vocabulary into broad categories as we have done in Section 5, it is not possible to do so at the teacher level and properly distinguish different feedback styles (e.g. "encouraging feedback", "competence-related feedback"). Yet we attempt to disentangle the effect of being exposed to a teacher who is relatively more likely to use the "female vocabulary" (respectively "male

vocabulary") for a given level of gendered feedback.

For that purpose, we compute two additional GTV measures. For each teacher×class, we compute the share of correctly predicted females on the one hand (hereafter referred to as the GTV-female measure), and the share of correctly predicted males on the other hand (GTV-male measure). These two measures enable us to highlight different patterns in the vocabulary used by teachers. For example, a teacher for whom we predict very accurately his female students' gender and poorly that of his male students would have a high GTV-female and a low GTV-male. This means that this teacher is very likely to use the female-specific vocabulary for his female students while he would use the gender-neutral or the female-specific vocabulary for his male students.

Panel (a) of Figure F9 shows the distributions of the overall *leave-one-out* GTV and of the *leave-one-out* GTVs computed for male and female students separately. This graph shows that male students are more often correctly classified by our model given that, on average, 65 percent of male students' gender is correctly predicted against 61 percent for female students. Panel (b) plots the correlation between GTV-male and GTV-female and further shows that teachers for whom one gender is often correctly predicted have a substantially lower proportion of correctly classified observations for the other gender. A one standard deviation increase in GTV-male is associated with a 0.7 standard deviation decrease in the GTV-females measure. This suggests that teachers using the female-specific vocabulary for their female students are not systematically using the male-specific vocabulary for their male students, but rather use the female or gender-neutral one.

To measure whether the positive effect on math performance at *baccalauréat* is reinforced by teachers more likely to use the vocabulary associated to female or that associated to male students, we estimate Equation (6) augmented with GTV-male (respectively GTV-female) in addition to the overall GTV measure. Table 8 reports the estimation results. It shows that the positive effect on math performance documented for higher-GTV teachers is reinforced by teachers with a higher GTV-female, i.e., teachers who are more likely to use the vocabulary associated to females. For a given level of GTV, a one standard deviation increase in GTV-female is associated with an additional 0.6 percent of a standard deviation increase in math performance on average. Our results suggest that exposure to a higher GTV-female teacher matters only for female students, for whom the standardized math grade increases by an additional 0.9 percent of a standard deviation against a non-significant 0.3 percent for male students. On the other hand, exposure to higher GTV-male teachers lowers the positive effect on math performance and seems detrimental to female students for whom math performance is reduced by 0.8 percent

of a standard deviation while male students seem mostly unaffected.

Altogether, the investigation of the different mechanisms conducted in this section suggests that exposure to gendered feedback affects performance beyond other teacher features. These results highlight the importance of considering teachers' written feedback as an additional input in the education production function.

8 Conclusion

Using comprehensive administrative data on the universe of Grade 12 students' transcripts, this paper shows how student gender affects the vocabulary used by math teachers in their written feedback. To identify gendered patterns in teacher feedback, we rely on machine learning techniques to predict students' gender based on the words used by their teachers. The key findings of the paper are threefold. First, we find that equally able female and male students receive different feedback, and that the scope of gender differentiation in teachers' feedback is large: the words used by Grade 12 math teachers allow to correctly predict the sex of 63 percent of the students on average. As a point of comparison, this is only marginally lower than the words' predictive power of students' performance level, which is the upper bound of what we could expect given that written feedback's purpose is to assess performance. Second, the qualitative analysis of the best gender predictors reveals that teachers insist more on positive managerial aspects and encourage the efforts provided by their female students, while equally performing males are both more criticized for their unruly behavior and praised for their 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 short run. To do so, we compute a teacher-level measure of gendered teacher vocabulary (GTV), that we define as the share of students whose gender is correctly predicted. Exploiting the quasi-random assignment of teachers within high school conditional on elective courses, we relate GTV to students' educational outcomes. This third key finding is that exposure to a teacher with higher GTV increases math performance by between 1.6 percent and 6 percent of a standard deviation, with slightly larger effects for female students. We do not find significant or sizeable effects of gendered teacher feedback on college applications or enrollment decision in the following year. Given the rather limited effects on math performance, the absence of effects on college outcomes is consistent if one considers performance as the main mediator between GTV and college outcomes. However, one could have expected GTV to influence college outcomes by

modifying aspirations and self-confidence. The absence of such effects suggests that this is not the case.

The magnitude of the effect of exposure to gendered feedback is in the lower bound of what we and other papers find regarding other inputs of the education production function. If we consider the effect of teacher quality on test scores, which constitutes an upper bound of the expected effects on student performance, Chetty et al. (2014) find that a one standard deviation increase in teacher value-added improves math test scores by 14 percent of a standard deviation, while we find estimates ranging between 16 percent and 20 percent in our setting. In the grading bias literature, Terrier (2020) finds for instance that having a teacher who is one standard deviation more biased against boys increases girls' progress in math by about 10 percent of a standard deviation. Carlana (2019) finds that exposure to teachers holding one standard deviation higher implicit stereotypes widens the gender gap in math by 4 percent of a standard deviation, which is more comparable to what we find.

Finally, we explore a range of mechanisms potentially driving the effects on math performance. First, we provide suggestive evidence that gendered feedback matters *per se*, by controlling for three teacher features that are likely to correlate with GTV: gender grading bias, personalized feedback and teacher quality. Second, in an attempt to open to black box of gendered feedback and understand which aspects of feedback potentially matter in determining student performance, we compare the effects of exposure to teachers that use more or less the vocabulary associated to female (respectively male) students, for a given level of gendered feedback. We find that the effect of gendered feedback on math performance for boys is the same regardless of the type of vocabulary used, whereas it is larger for girls exposed to teachers using the vocabulary associated to female students, i.e., teachers underlining the positive behavior and efforts.

The main take-away of our study is an awareness message as it highlights how feedback may be an effective pedagogical tool to improve students' performance. It opens avenues for future research. As 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 largest effects on performance. To go one step further, additional research using experimental settings where teacher feedback types randomly vary is needed.

References

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.

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.

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.

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

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, K., M. Ugalde-Araya, and B. Zafar, "A (Dynamic) Investigation of Stereotypes, Belief-Updating, and Behavior," *NBER Working Paper*, 2021, *29382.*

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.

Dupas, P., A. Sasser-Modestino, M. Niederle, and J. Wolfers, "Gender and the Dynamics of Economics Seminars," *NBER Working Paper*, 2021, *28494.*

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

Eble, Alex, Anjali Adukia, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz, "What We Teach About Race and Gender: Representation in Images and Text of Children's Books," *NBER Working Paper*, 2021, *29123.*

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.

Gallen, Y and M Wasserman, "Informed choices: Gender gaps in career advice," *Working paper*, 2021.

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

Good, C., A. Rattan, and C. Dweck, "Why Do Women Opt Out? Sense of Belonging and Women's Representation in Mathematics," *Journal of Personality and Social Psychology*, 2012, *102* (4), 700–717.

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," *Working Paper*, 2021.

Koenka, A. and E. Anderman, "Personalized Feedback as a Strategy for Improving Motivation and Performance Among Middle School Students," *Middle School Journal*, 2019, *50* (5), 15–22.

Koffi, M., "Innovative Ideas and Gender Inequality," *Job Market Paper*, 2020.

Landaud, Fanny, Eric Maurin, Barton Willage, and Willén Alexander, "Getting Lucky: The Long-Term Consequences of Exam Luck," *CESifo Working Papers*, 2022.

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.

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.

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*, 2021.

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.

Stepner, M., "VAM: Stata Module to Compute Teacher Value-Added Measures," *Statistical Software Components*, 2013, *S457711.*

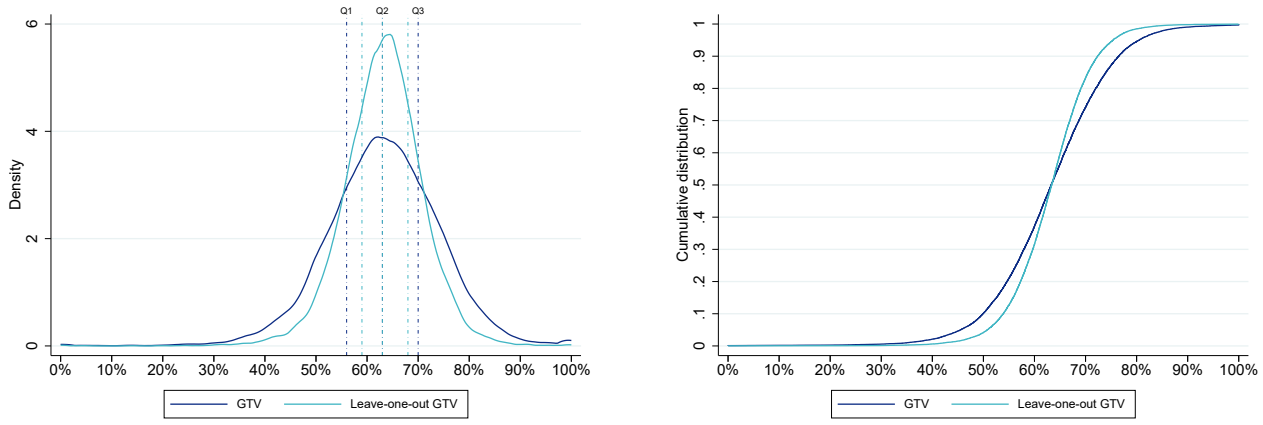
Terrier, C., "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement," *Economics of Education Review*, 2020, *77.*

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.

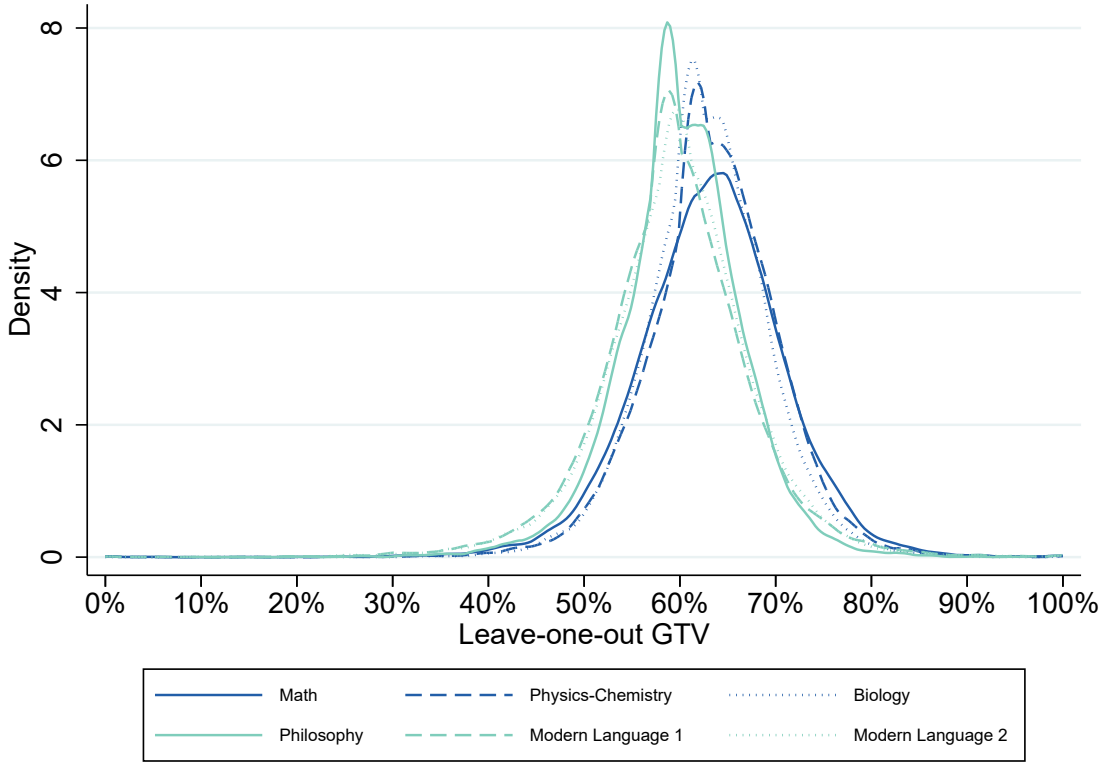
Figure 1 – Distribution of Math Teachers GTV and Leave-one-out GTV



(a) Density **(b)** Cumulative distribution

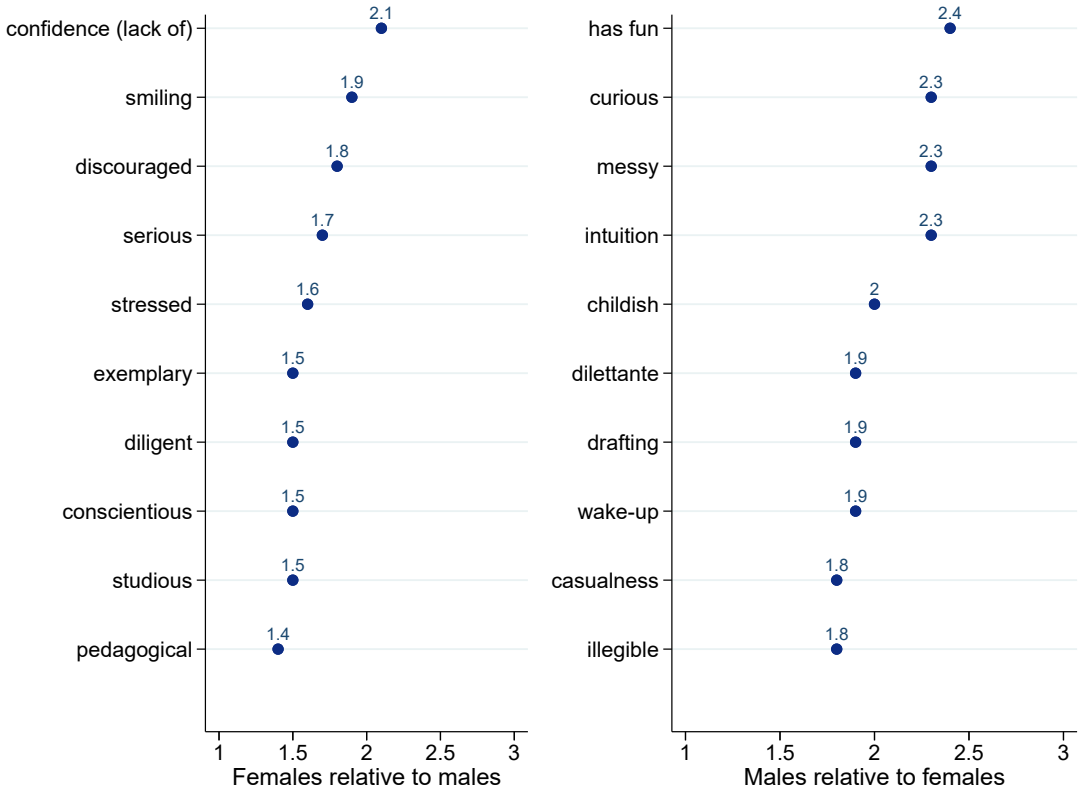
Notes: This figure shows the densities (Panel (a)) and cumulative distributions (Panel (b)) of the math teachers' GTV and leave-one-out GTV measures. The vertical lines in Panel (a) represent the first, second and third quartiles of the GTV distributions. Computations are based on administrative data from the French Ministry of higher education. The sample consists of Grade 12 math teachers teaching in high school × elective × year cells containing more than one math teacher.

Figure 2 – Distribution of Teachers' Leave-one-out GTV – By Core Subjects



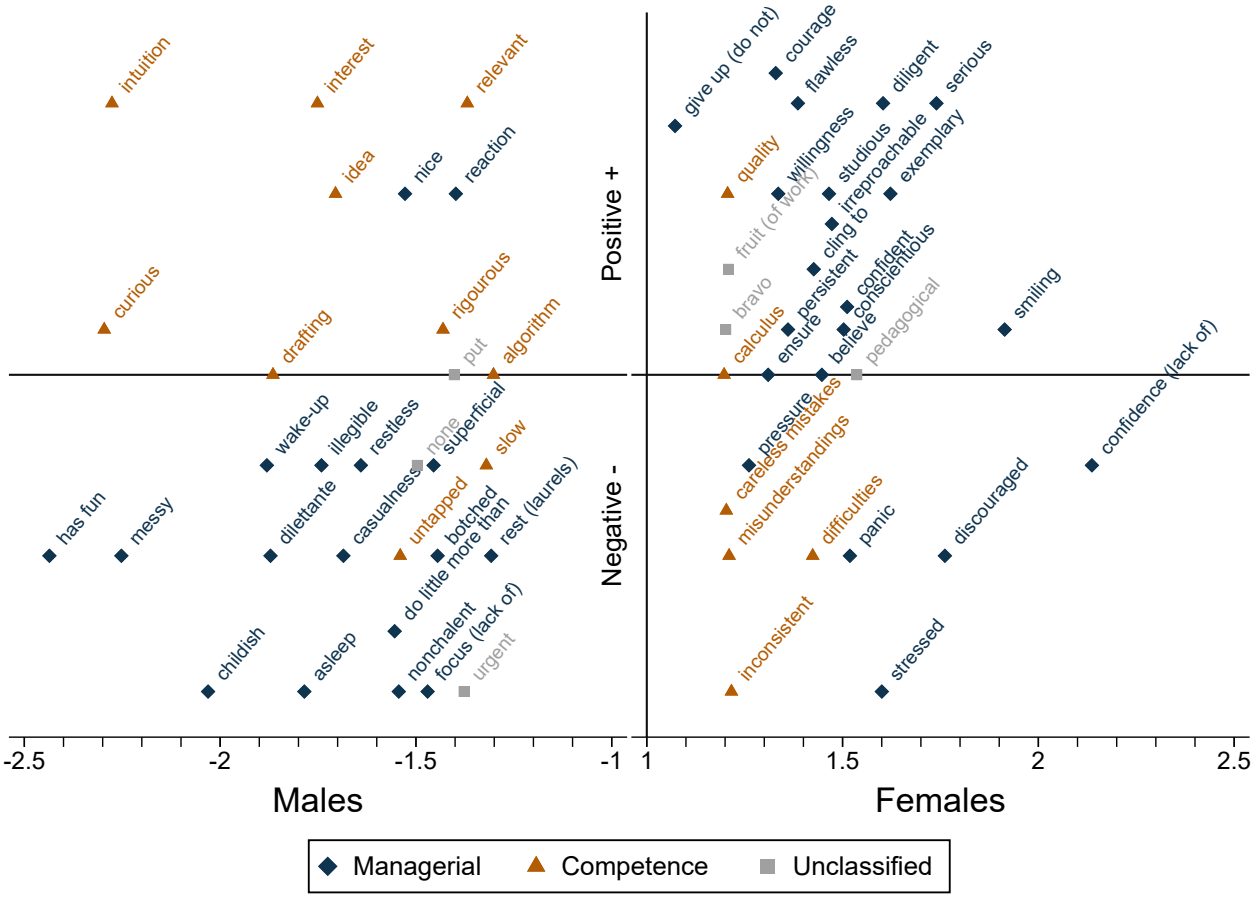
Notes: This figure shows the distributions of the math, physics, biology, philosophy and foreign language teachers' leave-one-out GTV measure, based on administrative data from the French Ministry of Education. The sample consists of Grade 12 teachers teaching in high school × elective × year cells containing more than one math teacher. Density distributions are all statistically different from each other at the 1 percent level as suggested by pairwise Kolmogorov-Smirnov tests for equality.

Figure 3 – 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. The estimation is realised on the universe of French Grade 12 science major students over the period 2012-2017.

Figure 4 – Classification of the Top 30 Gender Predictors



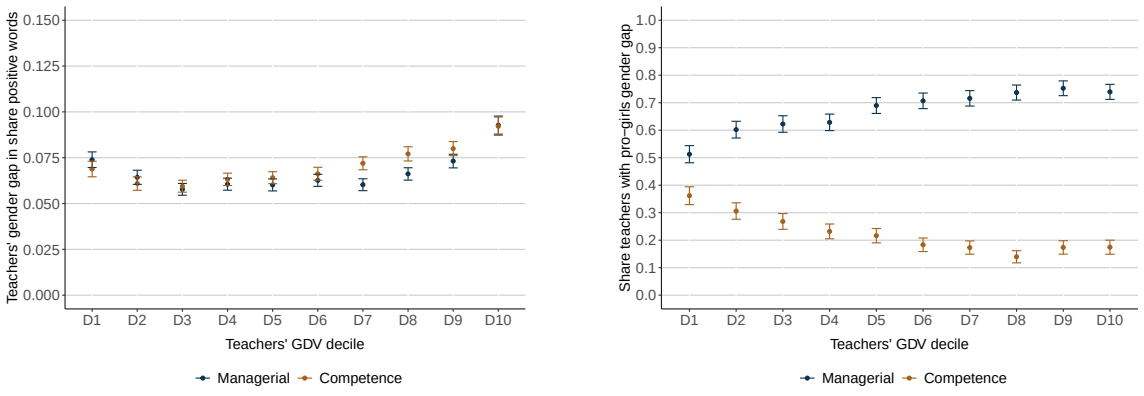
Notes: This figure classifies the top 30 female and male predictors of the model described by Equation (1) estimated using the vocabulary appearing in math teachers' feedback into positive vs. negative and managerial vs. competence categories. Ambiguous words (i.e. the ones used in both positive and 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. The estimation is realised on the universe of French Grade 12 science major students over the period 2012-2017.

Figure 5 – Gender Predictors' Type and Positiveness

(a) Positiveness conditional on feedback type **(b)** Feedback type conditional on positiveness

Notes: This figure used the classification of all the male and 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 managerial vs. competence 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 managerial and competence-related gender predictors conditional on positiveness, and Panel b shows the proportions of positive, neutral and negative gender predictors conditional on being competence-related or managerial.

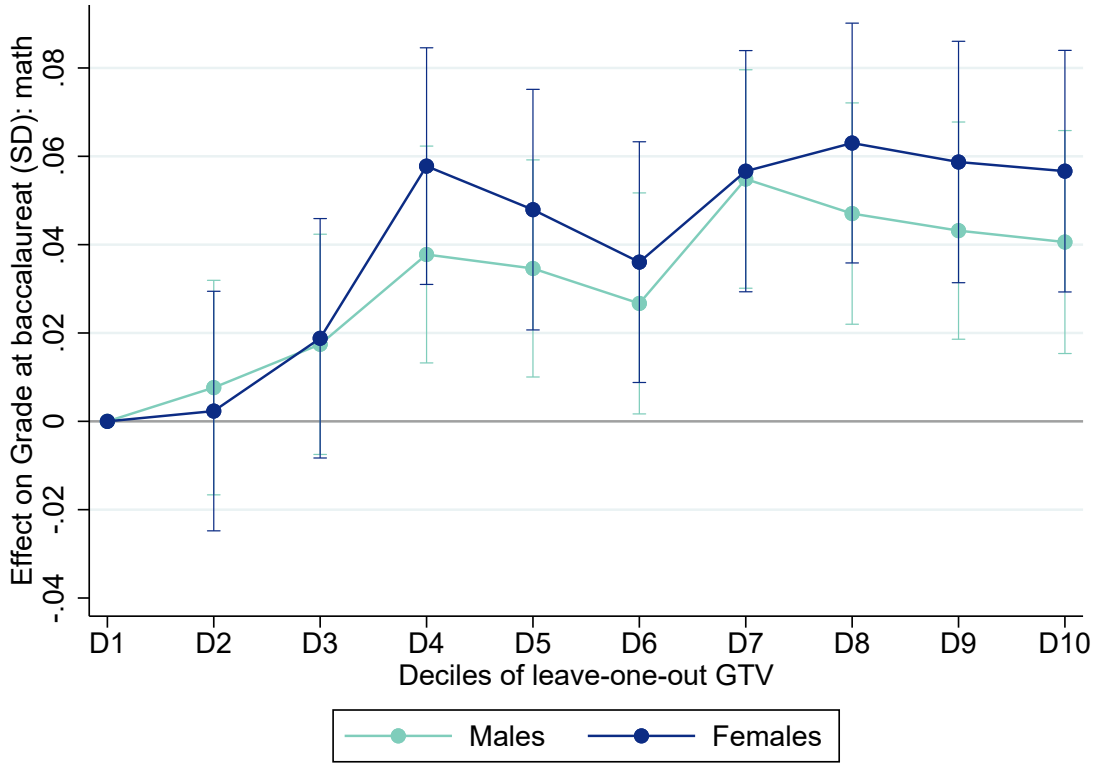
Figure 6 – Teachers' Gender Gap in the Share of Positive Words in Favor of Females by Deciles of GTV - In Absolute Value and Percentage



(a) Teacher's gender gap (absolute value) **(b)** Share of teachers with gender gap in favor of females

Notes: For each GTV decile, Panel (a) displays the average absolute value of Grade 12 teachers' gender gaps in the share of positive words appearing in their feedback, separately for competence vs. managerial related words. The GTV deciles are computed from the leave-one-out GTV. Panel (b) displays the share of teachers for whom the gender gap is in favor of female students, by GTV decile. The average values per decile are computed on the universe of math Grade 12 teachers for whom at least one GTV measure was estimated.

Figure 7 – 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. The coefficients are estimated using administrative data from the French Ministry of higher education over the period 2012-2017, and on the sample of Grade 12 science major students for whom the high school × elective × year cell contains more than one math teacher.

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

	2012	**2013**	**2014**	**2015**	**2016**	**2017**
Total nb. of G12 science major students	174,996	179,625	183,693	190,980	198,573	203,262
Nb. of obs. with missing transcript	90,328	79,233	54,256	42,445	33,999	28,500
% high school entirely missing:	95.6	92.6	91	85.6	78.2	68.1
High school < 2 classes	3,641	4,722	6,449	6,857	7,501	7,660
Teachers < 2 classes	14,191	5,775	5,903	5,917	8,900	32,030
Obs. in the analytical sample	66,836	89,895	117,085	135,761	148,173	135,072
(in %)	(38.19)	(50.05)	(63.74)	(71.09)	(74.62)	(66.45)

Notes: This table reports the number of Grade 12 science major applicants on APB 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 student (N= 691,234)	0.47	0.00	1.00
Age (years) (N= 691,234)	18.09	18.12	18.06
Free lunch student (N= 691,200)	0.13	0.12	0.14
High SES (N= 691,234)	0.43	0.44	0.41
Medium-high SES (N= 691,234)	0.16	0.16	0.16
Medium-low SES (N= 691,234)	0.24	0.24	0.25
Low SES (N= 691,234)	0.17	0.16	0.18
Education: past academic performance			
Rank at DNB: math (N= 655,152)	50.29	52.19	48.14
Rank at DNB: French (N= 655,121)	50.33	44.69	56.72
Rank at *baccalauréat*: French (written) (N= 659,484)	49.99	45.01	55.61
Rank at *baccalauréat*: French (oral) (N= 659,447)	49.79	45.70	54.40
Education: G12 elective course choice			
Maths elective (N= 623,112)	0.23	0.27	0.19
Physics-chemistry elective (N= 623,112)	0.26	0.27	0.25
Earth & life science elective (N= 623,112)	0.37	0.26	0.50
Engineering & computer science elective (N= 623,112)	0.13	0.20	0.06
Nb. of observations	691,234	369,056	322,178

Notes: This table shows descriptive statistics for Grade 12 science major students on the whole analytical sample, and separately for males and females. The number of non-missing observations is reported in parentheses.

Table 3 – Math Teachers' Summary Statistics

	Mean	**S.d**
Share of head teacher at least once (N= 6,751)	0.53	0.50
Male math teacher (N= 6,718)	0.58	0.49
Number of teacher observations (N= 6,770)	3.70	1.65
Average number of classes per year (N= 6,770)	1.09	0.26
Average number of students per class (N= 6,770)	28.04	5.20
Average feedback length (N= 6,754)	12.51	4.21
Nb. of teachers	6,770	

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. The number of non-missing observations is reported in parentheses.

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

	Dep. var: leave-one-out GTV		
	Coeff.	**S.e**	**p-value**
Female student	−0.0028	0.0022	0.2015
Age (years)	−0.0024	0.0020	0.2315
Scolarship student	0.0025	0.0028	0.3791
Foreign student	0.0126*	0.0068	0.0643
High SES	−0.0149	0.1457	0.9186
Medium-high SES	−0.0183	0.1457	0.8999
Medium-low SES	−0.0190	0.1457	0.8962
Low SES	−0.0181	0.1458	0.9013
Rank at DNB: math	−0.0000	0.0000	0.1782
Rank at DNB: French	0.0000	0.0000	0.2321
Rank at Baccalaureat: French (written)	0.0001**	0.0000	0.0189
Rank at Baccalaureat: French (oral)	0.0000	0.0000	0.4763
High school×elective×year FE	Yes		
Nb. of observations	573,025		

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

	Nb. of nonmissing p-values	Nb. of significant p-values at 5%	Share of significant at 5%	Share of significant at 1%
Female student	21,940	2,459	11.21	3.40
Age (years)	19,797	1,608	8.12	2.59
Free lunch student	19,162	987	5.15	1.28
Foreign student	8,084	333	4.12	1.27
High SES	22,140	1,482	6.69	1.41
Medium-high SES	20,839	941	4.52	0.86
Medium-low SES	21,925	1,141	5.20	0.93
Low SES	20,398	1,118	5.48	1.20
Rank at DNB: math	22,483	1,381	6.14	1.18
Rank at DNB: French	22,484	1,523	6.77	1.39
Rank at baccalaureat: French (written)	22,486	1,665	7.40	1.58
Rank at baccalaureat: French (oral)	22,482	1,591	7.08	1.42

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' Math Performance

	All	**Boys**	**Girls**
	(1)	(2)	(3)
Academic performance			
Grade at *Baccalauréat* (SD): math	0.0164***	0.0137***	0.0208***
	(0.0026)	(0.0030)	(0.0033)
Type of programs ranked first in the ROL			
All STEM tracks	−0.0026***	−0.0054***	0.0006
	(0.0009)	(0.0012)	(0.0012)
Selective STEM	−0.0011	−0.0039***	0.0020**
	(0.0008)	(0.0011)	(0.0009)
University STEM	−0.0012**	−0.0015**	−0.0007
	(0.0005)	(0.0007)	(0.0007)
Vocational STEM	−0.0003	0.0002	−0.0007
	(0.0005)	(0.0008)	(0.0006)
Matriculation in the following year			
All STEM	−0.0022**	−0.0045***	0.0004
	(0.0009)	(0.0012)	(0.0011)
Selective STEM	−0.0019***	−0.0041***	0.0003
	(0.0007)	(0.0010)	(0.0007)
University STEM	−0.0003	−0.0005	0.0000
	(0.0007)	(0.0011)	(0.0009)
Vocational STEM	0.0001	0.0002	−0.0001
	(0.0002)	(0.0003)	(0.0002)
Nb. of observations	717,578	383,350	334,228

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. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.01; **: p-value < 0.05, *: p-value < 0.1.

Table 7 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Math Performance - Mechanisms

	All	**Boys**	**Girls**
	(1)	(2)	(3)
Panel a. Teacher Grading Bias (SD)			
Grade at baccalaureat (SD): math			
Coeff. on GTV	0.0181***	0.0151***	0.0224***
	(0.0027)	(0.0031)	(0.0034)
Coeff. on mechanism	0.0123***	0.0135***	0.0117***
	(0.0028)	(0.0032)	(0.0037)
Panel b. Teacher Feedback Personalization (SD)			
Grade at baccalaureat (SD): math			
Coeff. on GTV	0.0163***	0.0136***	0.0208***
	(0.0026)	(0.0029)	(0.0033)
Coeff. on mechanism	0.0308***	0.0264***	0.0340***
	(0.0032)	(0.0035)	(0.0039)
Panel c. Teacher Value Added (SD)			
Grade at baccalaureat (SD): math			
Coeff. on GTV	0.0102***	0.0076***	0.0142***
	(0.0020)	(0.0026)	(0.0028)
Coeff. on mechanism	0.1830***	0.1660***	0.1997***
	(0.0043)	(0.0050)	(0.0053)
Nb. of observations	717,578	383,350	334,228

Notes: This table reports the coefficients on the standardized *leave-one-out* GTV obtained from the estimation of Equation (6) (row *Coeff. on GTV*). Each row 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* exam. The regression further controls for the standardized teacher grading bias (Panel a.), for the standardized measure of teacher feedback personalization (Panel b.), and for the standardized teacher value-added (Panel c.). 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 < 0.01; **: p-value < 0.05, *: p-value < 0.1.

Table 8 – Effects of Exposure to Male-Specific or Female-Specific Vocabulary on Students' Math Performance

	All (1)	**Boys** (2)	**Girls** (3)
Panel a. Higher Exposure to Male-Specific Vocabulary			
Grade at baccalaureat (SD): math			
Coeff. on GTV	0.0184***	0.0147***	0.0240***
	(0.0029)	(0.0033)	(0.0037)
Coeff. on GTV Male	−0.0047*	−0.0021	−0.0079**
	(0.0029)	(0.0032)	(0.0036)
Panel b. Higher Exposure to Female-Specific Vocabulary			
Grade at baccalaureat (SD): math			
Coeff. on GTV	0.0134***	0.0124***	0.0157***
	(0.0031)	(0.0035)	(0.0039)
Coeff. on GTV Female	0.0058*	0.0030	0.0092**
	(0.0032)	(0.0036)	(0.0041)
Nb. of observations	717,578	383,350	334,228

Notes: Each panel reports the coefficients on the standardized *leave-one-out* GTV obtained from the estimation of Equation (6) augmented with GTV-male (Panel a) or GTV-female (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. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.01; **: p-value < 0.05, *: p-value < 0.1.

Appendix to

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

Pauline Charousset, Marion Monnet

July 2022

List of Appendices

A Measuring Gendered Teacher Vocabulary (GTV): Details of the Estimation Procedure

This appendix provides the details on the practical implementation for the different steps of the gendered teacher vocabulary (GTV) estimation procedure developped in Section 4.

A.1 Textual Data Preparation

The students' academic records consist of a corpus of *documents*, where a *document* corresponds to the feedback that a teacher gave 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. In order 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 all punctuation signs, but keep track of the position of full stops in order to identify the different sentences that composed the original text. We get rid of all first names (that are identified based on 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. In order 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 one if the word appears in the document, and zero otherwise. In the text analysis literature, groups of words are commonly denoted *ngrams*, were n corresponds to the number of words in the considered group of words. In our analysis, we choose to use as regressors *unigrams*, i.e., tokens composed of only one word.

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 predictors of students' gender. 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 maximize the penalized log-likelihood function, where λ is the regularization parameter:

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

The $\hat{\alpha}$ estimates are then used to predict students' gender. The GTV measure is computed based on those predictions, and is defined as the proportion of students for whom the model correctly predicts their gender, separately for each teacher. 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 from each quartile of prior math performance. We define quartiles of prior math performance based on the math grade obtained on the DNB exam. 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 tokens for lengthy feedback, defined as the ones with an above-median length. For such feedback we randomly select 12 tokens, which is the median number.

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:

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

[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.

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 teacher GTV measure defined by Equation (A.3) could capture some unobserved-class specific gender differences. To rule out this concern, we also compute the *leave-one-out* teacher GTV as the average GTV over all the other classes taught except the current one:

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

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.

B Additional Results on Feedback Classification

B.1 Classification of the Top 100 Gender Predictors

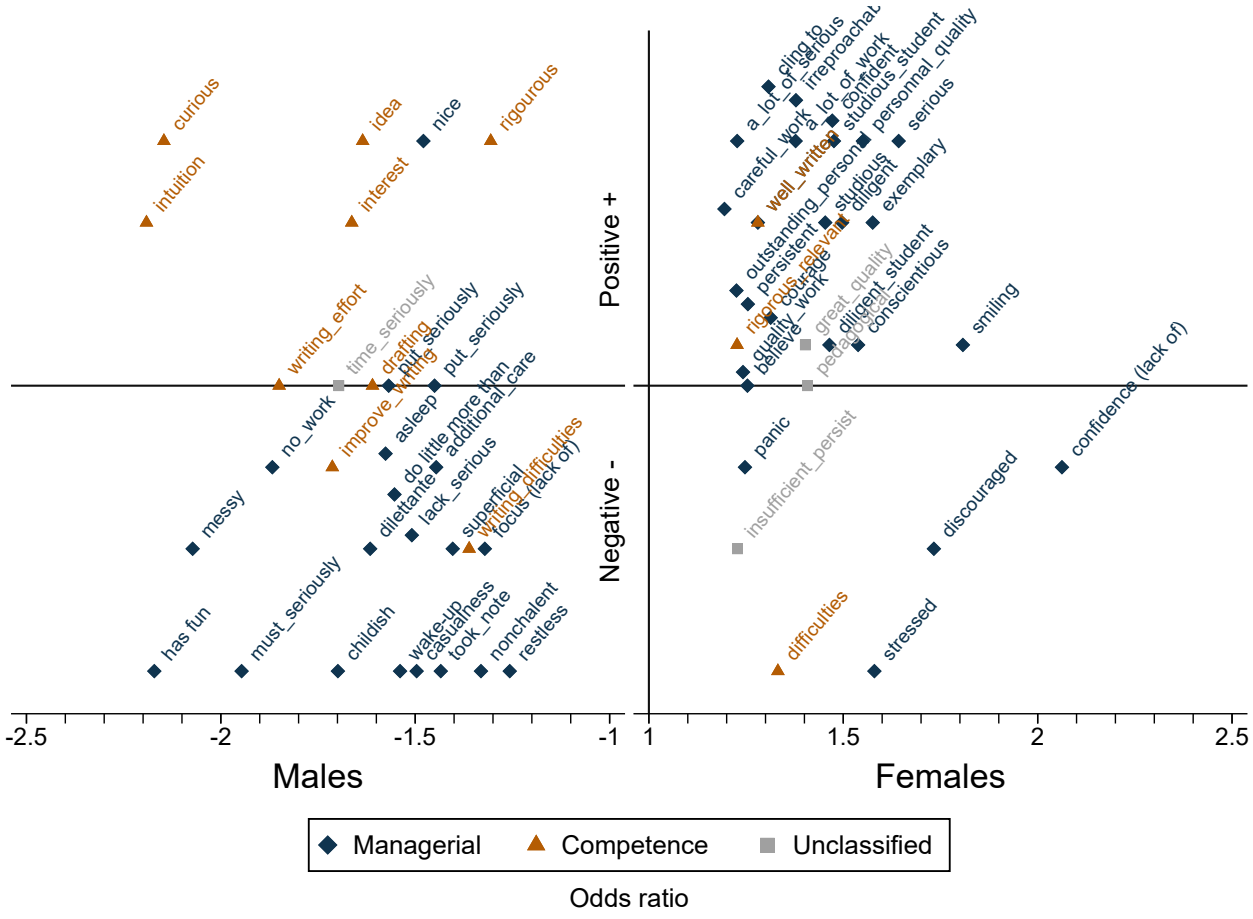
Table B2 – Top 100 Predictors' Classification - Female

	Positive	**Negative**	**Neutral**
Competence-related	**4 tokens:** accurate, autonomous, master, quality	**6 tokens:** careless mistakes, difficulties, inconsistent, mishap, mistake, misunderstandings	**11 tokens:** appropriate, assessment, calculus, elementary, litteral, method, methodological, question, read, test, theoretical
Managerial	**29 tokens**: abnegation, cling to, confident, conscientious, courage, deserve, determined, diligent, discrete, efficient, encourage, exemplary, fight, flawless, give up (do not), irreproachable, keep doing, persevere, persistent, pleasant, reassure, reward, serious, smiling, steady, studious, tenacious, voluntary, willingness	**12 tokens:** chattering, concern, confidence (lack of), discouraged, hesitate, panic, pressure, shy, stressed, suffer, unassuming, worry	**6 tokens:** believe, check, dare, ensure, intervene, pursue
Unclassified	**5 tokens:** bravo, congratulations, fruit (of work), pays off, reduce	**4 tokens:** decline, decrease, fragile, too low	**23 tokens:** (undefined), a lot, allow, also, benchmark, big, complete, contribute, despite, from now on, furthermore, help, illustrate, know, link, long, other, pedagogical, point, pupil, target, valid

Table B3 – Top 100 Predictors' Classification - Male

	Positive	**Negative**	**Neutral**
Competence-related	**14 tokens:** ambition, aptitude, capability, capacities, curious, gifted, idea, interest, intuition, passion, potential, relevant, rigourous, scientific	**2 tokens:** slow, untapped	**15 tokens:** algorithm, argument, computing, contest, culture, drafting, expression (oral/written), guidelines, homework, passage, reflex, word, write, writing, written
Managerial	**5 tokens:** consciousness, detailed, nice, reaction, worker	**26 tokens:** asleep, botched, care (lack of), casualness, childish, dilettante, disorganized, do little more than, focus (lack of), has fun, illegible, immature, inexistant, messy, minimal, nonchalent, rest (laurels), restless, scattered, shake up, skim through, superficial, troublesome, lets himself live, wake-up, waste	**7 tokens:** behave, exploit, in-depth, intensify, intervene, justify, work
Unclassified	**3 tokens**: best, easy, sufficient	**8 tokens**: excessive, insufficient, minimum, none, perfectible, shame, sufficient, urgent	**21 tokens:** (undefined), a while, advice, confirm, could, day, decide, expected, handed in, imposed, invite, lives, mature, personal, put, radical, time, took, wait

Figure B1 – Classification of the Top 30 Gender Predictors (with bigrams)



Notes: This figure classifies the top 30 female and the top 30 male 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 best predictors are classified into positive vs. negative and managerial vs. competence 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. The estimation is realised on the universe of French Grade 12 science major students over the period 2012-2017.

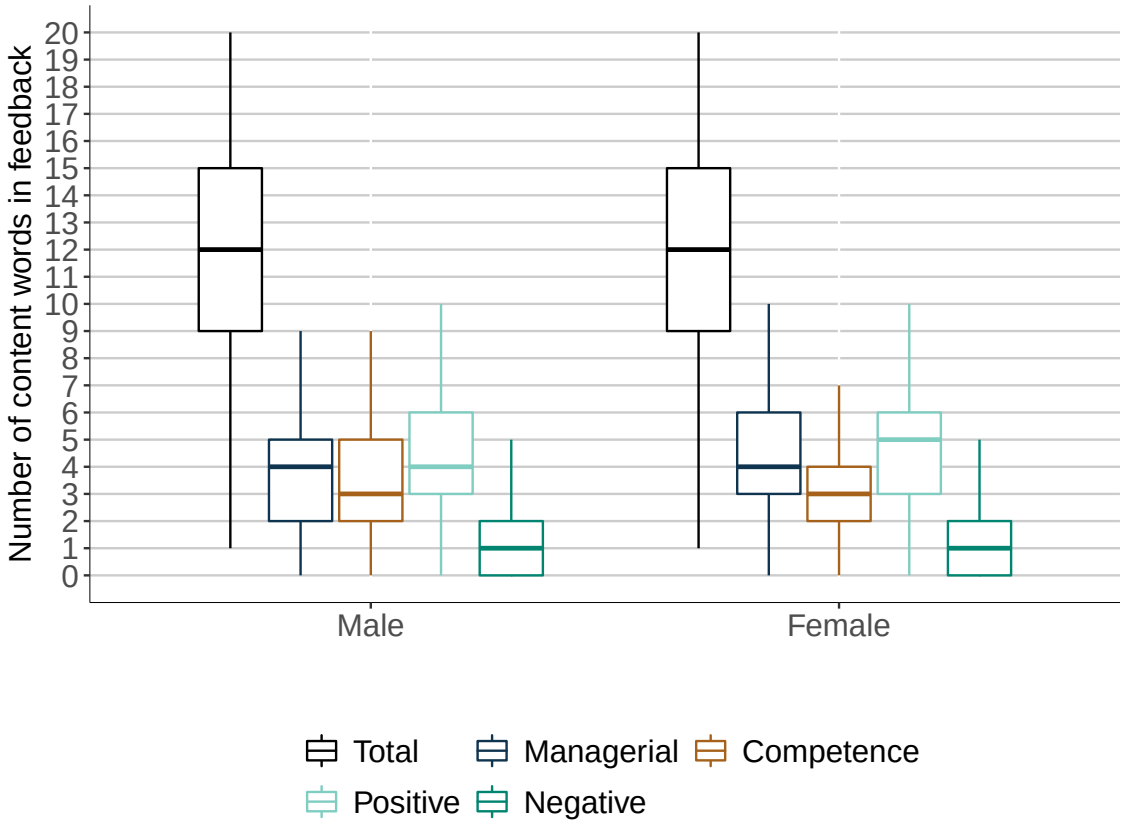
B.2 Classification of the Top 30 Gender Predictors - Bigrams

B.3 Descriptive Statistics and Classificiation on General Math Feedback

This appendix section aims at providing a broad picture of what a raw math feedback looks like on average for a Grade 12 science major student. We provide statistics on the distribution of word counts of math feedback overall and by type of feedback.

Panel (a) of Figure B2 displays basic summary statistics on the distribution of the number of content words appearing in the math feedback received by male and female students separately. Female and male students tend to receive feedback of the same length with a medium number of content words equal to 12. These summary statistics are then broken down according to the dimensions mentioned in Section 5.2: managerial vs. competence-related feedback and positive vs. negative feedback. The summary statistics along the positive vs. negative dimensions show that 50 percent of females get 5 positive words or more against 4 for males. The managerial and competence-related dimensions also highlight different gender patterns, at the top of the

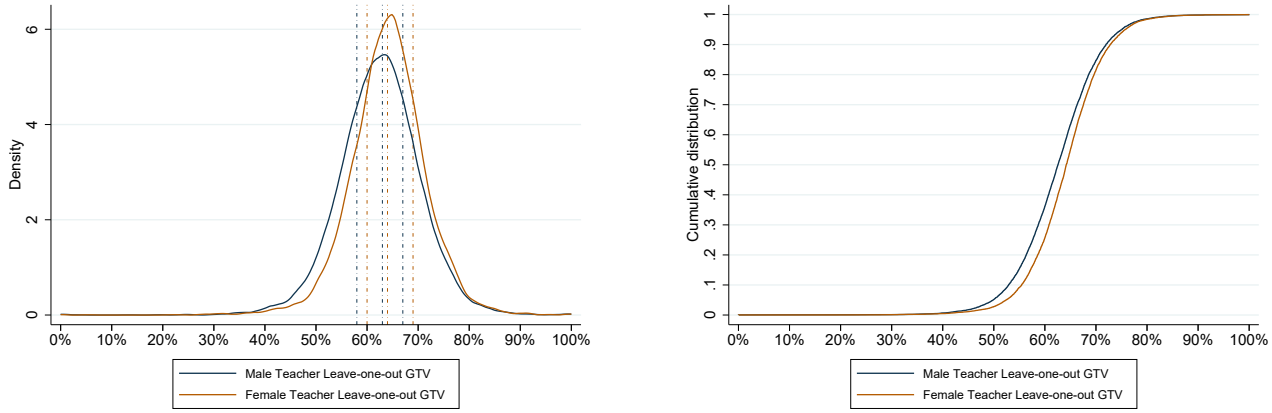
Figure B2 – Math Feedback - Distribution of Word Counts



Notes: This graph displays basic summary statistics on Grade 12 science major female and male students' distributions of feedback length in math, based on administrative data from the French Ministry of higher education. Each box displays the first and third quartile values as well as the median values. The segments cover the feedback length values that range between the first and third quartile values +/- 1.5 × IQR, where IQR denotes the interquartile range.

distribution only. The median feedback addressed to male and female students contains 3 competence-related word and 4 managerial-related words. However, 25 percent of female students receive more than 6 managerial words against 5 for males, and the reverse holds for compentence-related feedback: 25 percent of males get at least 5 such words against 4 for females. Note that, contrary to our statistical model, such differences may reflect actual differences in students' characteristics, such as differences in prior math ability between male and female students.

Figure C3 – Distribution of Math Teachers' Leave-one-out GTV – By Teacher Gender



(a) Density **(b)** Cumulative distribution

Notes: This figure shows the densities (Panel (a)) and cumulative distributions (Panel (b)) of male and female math teachers' leave-one-out GTV measures. The vertical lines in Panel (a) represent the first, second and third quartiles of the GTV distributions. Computations are based on administrative data from the French Ministry of higher education. The sample consists of Grade 12 math teachers teaching in high school × elective × year cells containing more than one math teacher.

C Statistics by Teacher Gender

Table C4 – Share of Male Teachers by Core Subjects

Subject	Share	N	% non-miss
Math	0.58	7,121	0.93
Physics-Chemistry	0.57	7,764	0.93
Biology	0.37	6,698	0.92
Philosophy	0.62	7,412	0.95
Modern language 1	0.20	17,574	0.88
Modern language 2	0.19	22,625	0.83

Notes: The table reports the share of male teachers in the six core subjects taught in Grade 12 science major.

D Assessing the Randomness of Missing Grade Transcripts

Table D5 – Balancing Test: High Schools with All Missing Grade Transcripts

	Dep. var: Grade transcripts all missing in high school		
	Coeff.	**S.e**	**p-value**
Female student	-0.1162^{***}	0.0362	0.0013
Age (years)	0.1863^{***}	0.0141	0.0000
Free lunch student	-0.3860^{***}	0.0485	0.0000
Foreign student	0.0057	0.0871	0.9478
High SES	0.0116	0.0421	0.7839
Medium-high SES	-0.3589^{***}	0.0691	0.0000
Medium-low SES	-0.1226^{**}	0.0535	0.0219
Rank at DNB maths	0.0024	0.0021	0.2693
Rank at DNB math (females)	0.0012	0.0010	0.1956
Rank at DNB math (males)	-0.0039^{***}	0.0013	0.0023
Nb. of observations	12,864		

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 Robustness Checks and Additional Results

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

Table E6 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Performance in Other Core Subjects

	All	**Boys**	**Girls**
	(1)	(2)	(3)
Academic performance			
Grade at *Baccalauréat* (SD): physics	−0.0027	−0.0046*	−0.0009
	(0.0022)	(0.0027)	(0.0029)
Grade at *Baccalauréat* (SD): biology	−0.0032	−0.0080***	0.0035
	(0.0024)	(0.0030)	(0.0031)
Grade at *Baccalauréat* (SD): philosophy	0.0007	−0.0006	0.0009
	(0.0025)	(0.0029)	(0.0031)
Nb. of observations	717,578	383,350	334,228

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. Standard errors are clustered at the teacher level and are reported in parentheses. ***: p-value < 0.01; **: p-value < 0.05, *: p-value < 0.1.

E.2 Robustness Checks

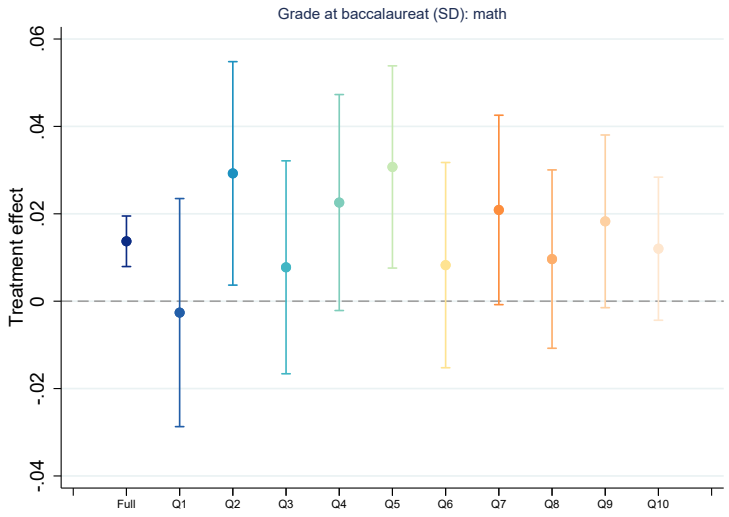
Table E7 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Educational Outcomes - Robustness Checks

	Boys			Girls		
	Bsl X (1)	*Share girls* (2)	*GTV other* (3)	*Bsl X* (4)	*Share girls* (5)	*GTV other* (6)
Academic performance						
Grade at *baccalauréat* (SD): math	0.0123***	0.0137***	0.0137***	0.0182***	0.0207***	0.0210***
	(0.0027	(0.0030	(0.0030)	(0.0029)	(0.0033)	(0.0033)
Type of STEM programs ranked first in the ROL						
All STEM tracks	−0.0058***	−0.0053***	−0.0053***	0.0005	0.0007	0.0007
	(0.0013	(0.0012	(0.0012)	(0.0012)	(0.0012)	(0.0012)
Selective STEM	−0.0045***	−0.0039***	−0.0039***	0.0015*	0.0020**	0.0020**
	(0.0012	(0.0011	(0.0011)	(0.0009)	(0.0009)	(0.0009)
University STEM	−0.0011	−0.0015**	−0.0015**	−0.0007	−0.0007	−0.0007
	(0.0007	(0.0007	(0.0007)	(0.0008)	(0.0007)	(0.0007)
Vocational STEM	−0.0000	0.0002	0.0003	−0.0005	−0.0007	−0.0007
	(0.0009	(0.0008	(0.0009)	(0.0006)	(0.0006)	(0.0006)
Matriculation in the following year						
All STEM	−0.0049***	−0.0045***	−0.0045***	0.0005	0.0005	0.0005
	(0.0013	(0.0012	(0.0012)	(0.0011)	(0.0011)	(0.0011)
Selective STEM	−0.0046***	−0.0041***	−0.0042***	−0.0002	0.0003	0.0003
	(0.0010	(0.0010	(0.0010)	(0.0007)	(0.0007)	(0.0007)
University STEM	−0.0004	−0.0005	−0.0004	0.0005	0.0000	0.0001
	(0.0011	(0.0010	(0.0011)	(0.0010)	(0.0009)	(0.0009)
Vocational STEM	0.0002	0.0002	0.0002	−0.0001	−0.0001	−0.0001
	(0.0003	(0.0003	(0.0003)	(0.0002)	(0.0002)	(0.0002)
Nb. of observations	717,578	383,350	334,228			

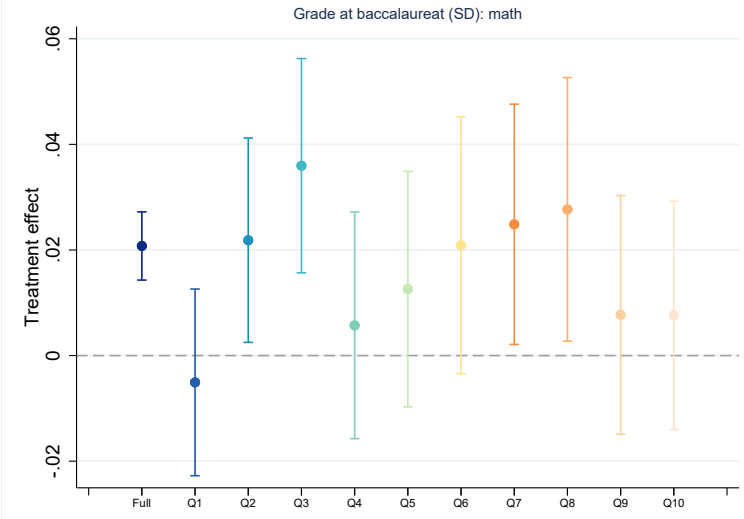
Notes: Each row reports the coefficients on the standardized *leave-one-out* teacher GTV obtained from the estimation of Equation (6) for the different outcomes listed on the first column. It is estimated on the whole sample and separately for Grade 12 male and female students. The regression includes high school×elective course×year fixed effects. 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 < 0.01; **: p-value < 0.05, *: p-value < 0.1.

E.3 Additional Results: Heterogeneity by Initial Math Performance

Figure E4 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Math Performance at *baccalauréat* - By Deciles of Initial Math Performance



(a) Males

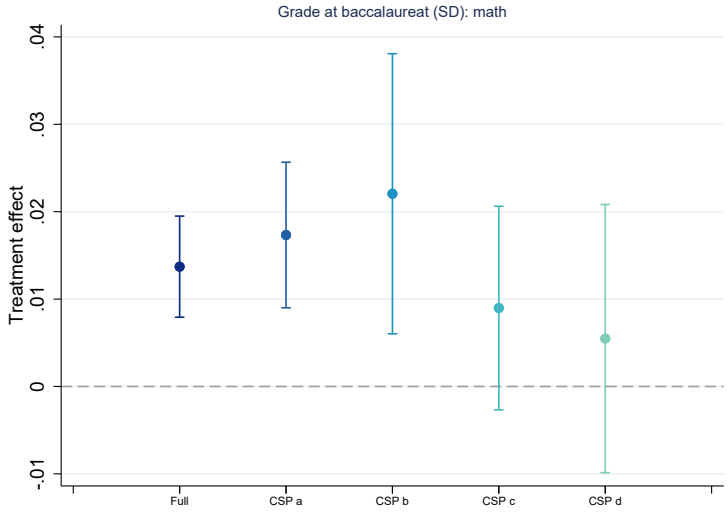


(b) Females

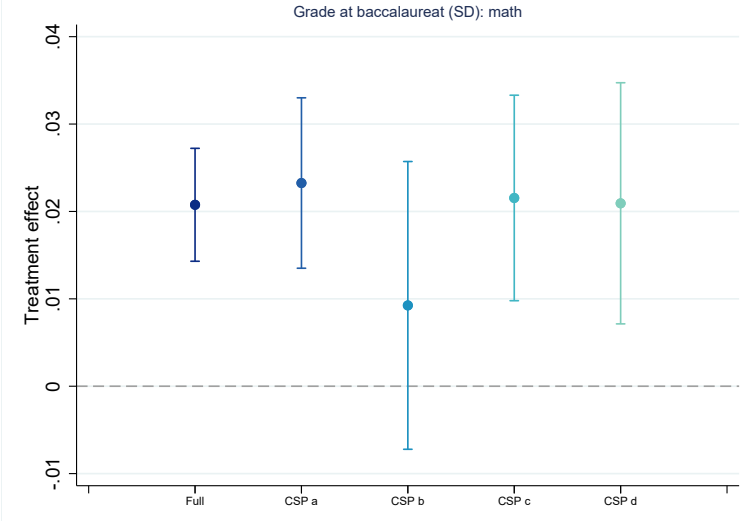
Notes: The figure reports the effect of a one standard deviation increase in leave-one-out GTV on students' standardized grade on the math *baccalauréat* exam separately by gender and by initial performance in math. Initial math performance is measured as deciles of percentile rank in math obtained at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

E.4 Additional Results: Heterogeneity by Social Background

Figure E5 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Math Performance at *baccalauréat*" - By Social Background



(a) Males



(b) Females

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.

F Mechanisms: Estimation Details and Complementary Results

F.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 at the continuous assessment as the non-blind score, and the (standardized) math grade obtained 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:

$$GB_{cjt} = \left(NB_{cjt}^{males} - NB_{cjt}^{females}\right) - \left(B_{cjt}^{males} - B_{cjt}^{females}\right)$$

The grading bias assigned to class c is actually the average bias observed in any other classes taught by the same teacher except class c itself, 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. We see that 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 F8 – Maths grades during G12 and at Baccalaureat exam - By students' and teachers' gender

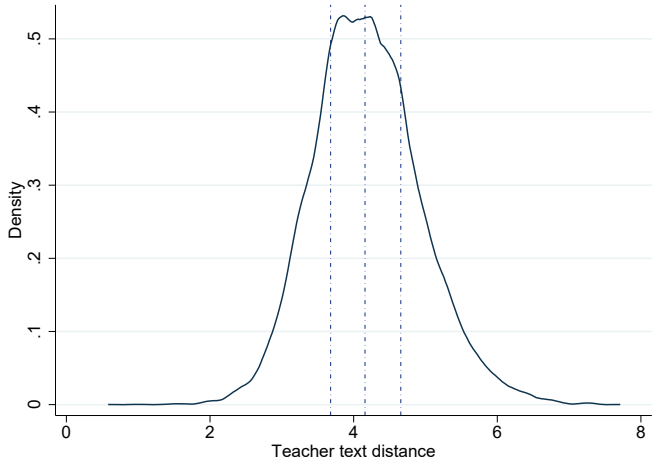
	Boys			**Girls**			**Teacher**
	G12 maths	Bac maths	Diff.	G12 maths	Bac maths	Diff.	**bias**
All teachers	−0.017	0.043	−0.061	0.020	−0.049	0.068	−0.129
Female teachers	−0.029	0.028	−0.057	0.033	−0.031	0.064	−0.121
Male teachers	−0.009	0.054	−0.063	0.010	−0.062	0.072	−0.135
N	364,769	344,131		319,552	306,618		

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.

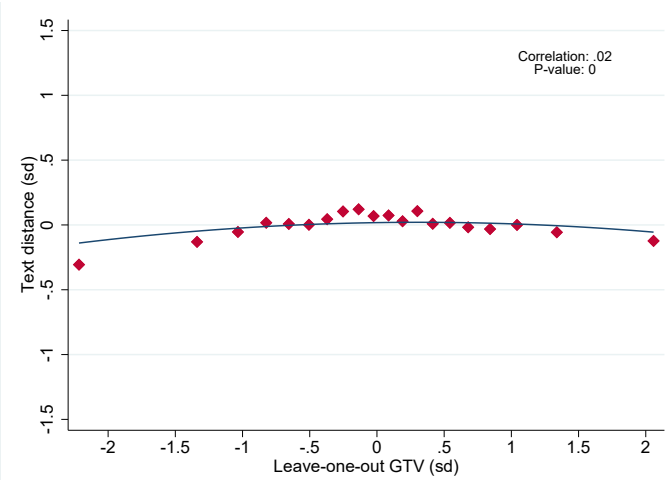
Figure F8 displays the correlation between our measure of GTV and the teacher grading bias. While statistically significant, the magnitude of the correlation between the two standardized measure is negligeable.

F.2 Teacher Feedback Personalization: Distribution and Correlation with GTV

Figure F6 – Distribution and Correlation of Teacher Feedback Personalization with GTV



(a) Density of text distance



(b) Correlation

Notes: The figure in Panel (a) displays the distribution of the teacher text distance, as measured by the euclidean distance between each of his written feedback. The vertical dotted lines represent the first, second and third quartile. Panel (b) shows the binned average of the teacher text distance measure for different values of standardised leave-one-out GTV. The line represents the quadratic fit. The correlation coefficients are obtained from the regression of the text distance on leave-one-out GTV.

F.3 Estimating the Teacher Value-Added

Teacher value-added is estimated using the three steps described in the Chetty et al. (2014) paper. The steps are implemented using the `vam` package developed by Stepner (2013). We detail these three steps below.

Step 1: Residualizing students test scores. We first regress students' test scores in year t, measured by the percentile rank obtained at the math *baccalauréat*, on a set of students' baseline covariates, controls for students' prior performance, previous year's class characteristics, and teachers fixed effects.

- *Students' baseline characteristics:* gender; free-lunch status; four dummies for students' SES background (low SES, medium-low SES, medium-high SES, high SES); a dummy equal to one if the student is a foreigner.

- *Students' prior performance:* It includes the math grade obtained during the Grade 11 continuous assessment, standardized by the mean and standard deviation of the class so that grades are comparable across classes. We also include its square and cube. We further control for the percentile rank at the math and French DNB national exam, as well as for the percentile rank at the French oral and written *baccalauréat* anticipated examinations.

- *Previous year's class characteristics:* It includes the average of all the students' characteristics listed above computed at the Grade 11 level, the class average at the math continuous assessment, the lowest and the highest math grade of the class.

After the regression, we predict students' test scores residuals adjusted for observables.[A.2] Finally, for each teacher's class in year t, we compute the average test score residual. This should be seen as a proxy for teachers quality in the class taught in year t.

Step 2: Regressing teachers' quality in year t on its lags and leads. We regress the average test score residuals of teachers in t on those average residuals in years $t-1, t-2, \ldots$ and $t+1, t+2, \ldots$. The OLS coefficients obtained from this regression tell us how strongly current teacher performance is related to its past and future performance, i.e., they are autocorrelation coefficients. These coefficients are also called *shrinkage* factors.

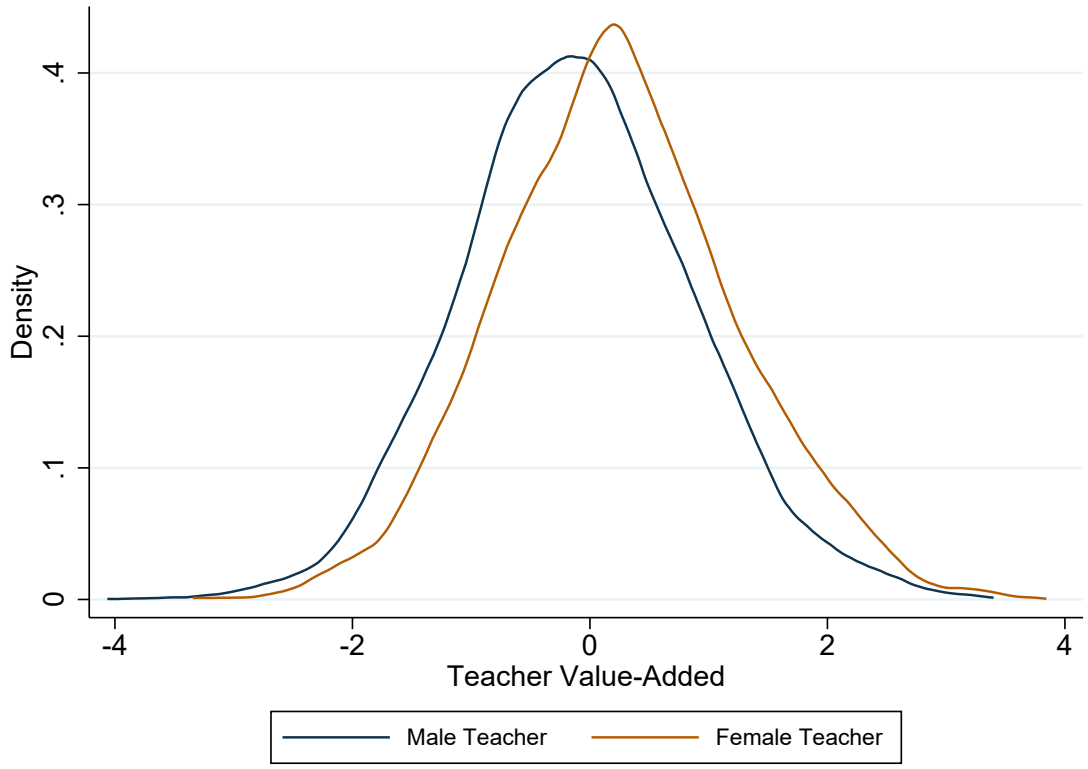
Step 3: Predicting teachers' quality. The final step consists in using the set of OLS coefficients from step 2 to *predict* teachers' quality. This predicted teacher quality is actually just a proxy for a teacher's true value-added and its reliability depends on the shrinkage factor, usually estimated to be around one-third (i.e., the true teacher value-added accounts for one-third of the residual variance).

The distribution of (standardized) teachers' predicted value-added is displayed in Figure F7.

Figure F8 displays the correlation between our measure of gendered teacher vocabulary and the teacher teacher value-added. Again, despite being statistically significant, the magnitude of the correlation between the two standardized measure is very low.

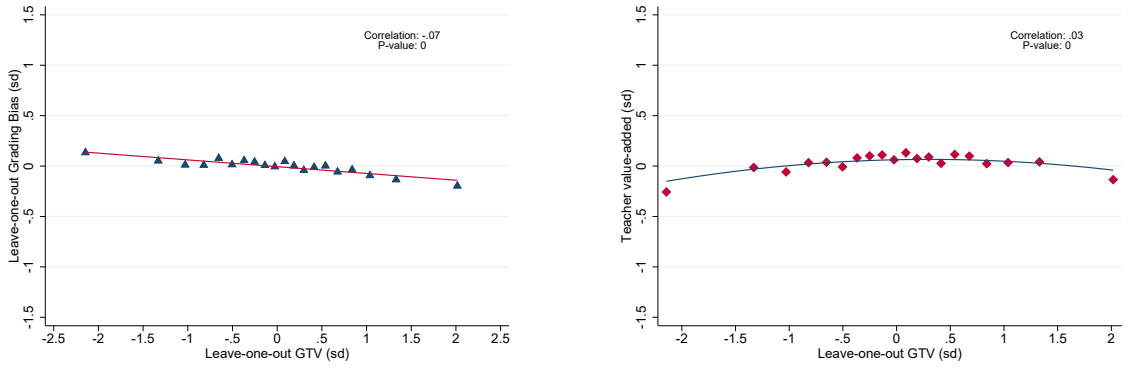
[A.2] Teacher fixed effects are included in the regression so that coefficients on other covariates are estimated only using the within teacher variation. Those fixed effects are then added back to the residuals.

Figure F7 – Distribution of Teachers' Predicted Value-Added



Notes: This graph plots the densities of math teachers' predicted value-added, separately for male and female teachers. The value-added estimates are obtained with the methodology described in Chetty et al. (2014) and implemented with the `vam` Stata package developped by Stepner (2013).

Figure F8 – Correlation Between GTV, Grading Bias and Teacher Quality

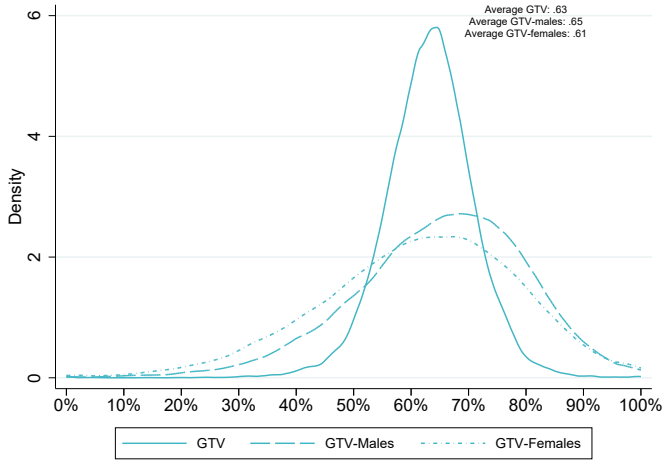


(a) Teacher GTV and Teacher Grading Bias **(b)** Teacher GTV and Teacher Value-Added

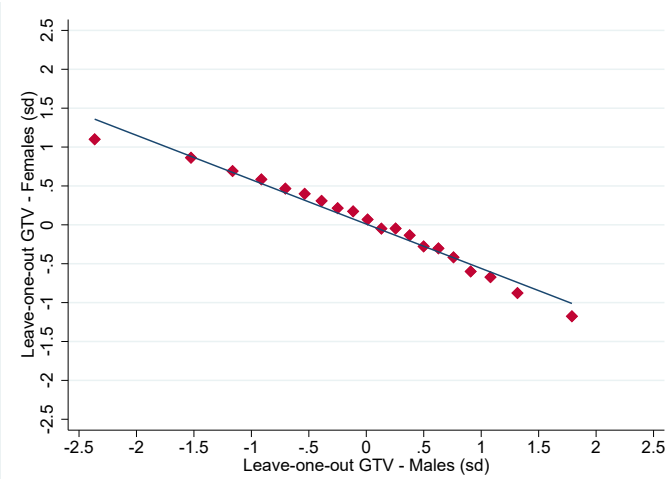
Notes: The figure shows the binned average of the teachers' leave-one-out grading bias (resp. value-added) standardised measures on the standardised teacher leave-one-out GTV. The line represents the linear fit in Panel (a) and the quadratic fit in Panel (b). The correlation coefficients are obtained from the regression of the grading bias (resp. value added) on teacher GTV. The sample consists of all Grade 12 math teachers for whom a leave-one out GTV measure, a leave-one-out grading bias measure and a value-added measure could be estimated.

F.4 GTV by Gender: Distribution and Correlation with GTV

Figure F9 – Distribution and Correlation of Leave-one-out GTV by Gender



(a) Density of leave-one-out GTV



(b) Correlation

Notes: Panel (a) of 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). Panel (b) shows binned averages of GTV-males and GTV-females and plots the fitted regression line.