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

Pauline Charousset

Marion Monnet

(Total word count : 11,650)

October 2022

Abstract

This paper studies how students' gender influences feedback from teachers, and how this in turn affects the students' performance. Using the written feedback to French high school students from 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. They highlight the positive behavior and encourage the efforts of their female students, while similarly-performing boys tend to be criticized for unruly behavior and praised for intellectual skills. To see how this relates to the students' subsequent educational outcomes, we then match these data with French national examinations results, these students' higher education applications and ultimate institution of enrollment. Exploiting the quasi-random allocation of teachers to classes, we estimate that having a teacher with feedback that is one standard deviation more gendered improves math performance by 1.6 percent of a standard deviation on average, but does not affect 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: Institute for Research in Education (Irédu) - University of Burgundy, 11 Esplanade Erasme, 21 000 Dijon, France (e-mail: marion.monnet@u-bourgogne.fr). We are particularly grateful to Anne Boring, Alex Eble, Julien Grenet and Clémentine Van Effenterre for numerous reviews and wise advice. This paper also benefited greatly from discussions and helpful comments from Elliott Ash, Asma Benhenda, Etienne Dagorn, Marc Gurgand, Elise Huillery, Sylvie Lambert, Arnaud Maurel, Dominique Meurs, Roland Rathelot, Michael Stepner, Camille Terrier, and all the participants in the IPP Reading group, Sciences Po Education Policies 2020 seminar, Paris School of Economics Labor Chair Seminar 2021, ETH Zurich Data Science and Economics Seminar 2021, French National Institute for Demographic Studies Seminar 2021, Journées de Microéonomie Appliquée 2021, IWAEE 2021, University of Rennes Young Economist Seminar 2021, ASSA/AEA meetings 2022, PSE Young Economists of Education Workshop 2022. We are especially grateful to the French Ministry of Higher Education for access to their database. Financial support from the French National Research Agency (Agence Nationale de la Recherche) through project ANR-17-CE28-0001, EUR grant ANR-17-EUR-0001 and Chaire Femmes et Sciences is gratefully acknowledged.

1 Introduction

Whether and how a student's gender influences the feedback he or she receives carries important implications for gender parity in education and in the labor market. Students' educational engagement and career choices depend significantly on the information about their ability that accumulates over their school career. This information is inherently noisy, in that it reflects not only intrinsic ability, but a whole range of other factors, such as assignment difficulty (Landaud et al., 2022), perseverance and effort (Alan et al., 2019), and others' perceptions of the student's performance (Sarsons, 2019). If teachers provide different kinds of feedback to male and female students, this could distort educational decisions, and possibly exacerbate gender differences in performance and career choices.

This paper provides an empirical test of whether student gender affects the feedback from math teachers in Grade 12 and examines how this relates to student performance and higher education enrollment choices. Using the entire set of Grade 12 student transcripts over the period 2012-2017 (available on the higher education applications platform), we analyze the written feedback from 6,770 math teachers to some 700,000 students in France. We find that student gender does in fact influence the feedback and show, further, that students whose teachers provide gendered feedback perform better on national examinations but make similar higher education applications and enrollment decisions. These findings are consistent with a direct effect of performance feedback on motivation, effort, and hence achievement, and a negligible impact on longer run outcomes such as self-perception and enrollment decisions.

The features of our data allow us to determine whether student gender influences feedback and to relate it to educational outcomes. Students' detailed school transcripts allow examination of teachers' written feedback – which is both highly relevant and highly informative to students – and thorough analysis of the particular wording used by math teachers in assessing male or female students' performance. We can then link the transcripts to national examination results, to assess how different feedback relates to student performance on these crucial, high-stakes tests. Third, the data permit investigation of how gendered feedback correlates with other teacher characteristics or teaching practices, such as 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, comparing the educational careers of students exposed to different degrees of gendered feedback.

Exploiting this rich source of information, our first contribution is the analysis of written feedback in the light of gender differences. We propose a synthetic measure of gendered teacher

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

Our first set of results provides evidence that math teachers do in fact provide differentiated feedback to equally good male and female students. The average GTV index is 63 percent – that is, on average, our model correctly predicts the gender of 63 percent of students. By way of comparison, this is only marginally lower than the words' predictive power for students' performance, which can be taken as a sort of upper bound, given that the express purpose of the feedback is precisely to assess performance. However, we also document considerable variation in the distribution of GTV, suggesting that students are exposed to significantly variable degrees of gendered feedback. To gauge whether the use of a gendered vocabulary is more prevalent in math, we replicate the analysis for five other subjects taken by our pool of Grade 12 students, finding that math teachers do in fact use a more gendered vocabulary than teachers in humanities.

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

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

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

intellectual skills.

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

Our third result is that having a teacher with a 1-standard-deviation higher GTV is associated with an average gain in performance on the math *baccalauréat* exam of 1.6 percent of a standard deviation; this effect is slightly greater for girls, but not significantly different from that of boys (2.1 percent vs. 1.4 percent of a standard deviation). The magnitude of the effect is admittedly moderate, but the impact on math performance on the *baccalauréat* is stronger for students who are exposed to teachers with an above-median GTV. Compared to students with teachers in the lowest decile of the GTV distribution, those with teachers in the fourth decile or above have math grades higher by up to 6 percent of a standard deviation. The effects on students' top-ranked program in their college application and on their enrollment outcomes are very small and mostly not significant.

Finally, we explore the possible mechanisms behind the effect of having a teacher with higher GTV on math performance. Although we cannot rule out some correlation of our measure of gendered feedback with other teacher characteristics, we try to exclude a series of alternative explanations. First, we show that our results are not driven by other teaching practices such as gender grading bias or feedback personalization as such. In line with Terrier (2020), we find evidence of teacher grading bias in math in favor of girls, but it is only weakly correlated with GTV. Our results are robust to controlling for this grading bias, and also robust to controlling for feedback personalization, proxied by a measure of text distance between the texts of 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 high GTV are also better teachers. Value-added and GTV turn out to be moderately correlated, and our results are channeled partly through teacher quality. Lastly, we investigate whether teachers' vocabulary affects male and female students' math performance differently. We find that the positive effect of exposure to gendered feedback is reinforced when teachers are more likely to use the vocabulary associated

with female students, i.e. emphasizing the positive behavior and effort. In keeping with the feedback and growth mindset literature, we find that females benefit more from the positive managerial feedback than from the negative behavioral feedback and intellectual praise, while for males no difference between the effects of the two types of feedback is statistically detectable.

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

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

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

levels.

Finally, our paper relates to the literature on the scope of gendered teacher behavior on student outcomes. Prior research provides evidence that teachers hold stereotyped beliefs about gender (Carlana, 2019), that they interact more with boys than with girls (Bassi et al., 2018), grade equally performing male and female students differently in their continuous assessment (Lavy and Sand, 2018; Terrier, 2020), and give them different career advice (Gallen and Wasserman, 2021). These gendered behaviors all have long-term consequences for academic schooling outcomes. To our knowledge, our paper is the first to provide direct evidence of gendered behavior for another teaching practice, namely written feedback, and to document the short-run effect on student performance and higher education enrollment decisions of having teachers who follow different feedback practices.

The rest of the paper is organized as follows. Section 2 gives some institutional background on French secondary education and on university admission. Section 3 describes the various data sources and provides some descriptive statistics on the population of Grade 12 students and their math teachers. Section 4 presents our empirical strategy, detailing the steps taken in constructing our gauge of gendered teacher vocabulary (GTV), and measuring its impact on students' outcomes. Section 5 analyzes the gendered vocabulary in detail, with some statistics on the distribution of our GTV measure. Section 6 shows the impact of having a relatively high-GTV teacher on academic performance, higher education preferences and enrollment outcomes in the year following graduation. Section 7 discusses potential mechanisms and Section 8 concludes.

2 Institutional Background

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

2.1 The French Secondary Education System

In France, secondary education consists of seven years of schooling: four years of middle school common to all students (*collège*, Grades 6 to 9), and three years of high school (*lycée*, Grades 10 to 12), which provide either vocational or general academic and technical training. Both the middle school and the high school curricula end with a national examination. At the end of middle school, students take the *Diplôme National du Brevet* (DNB), which tests their knowledge and skills in math, French and history and geography. At the end of Grade 11, high school

students take the preliminary *baccalauréat* examinations, which include oral and written tests in French, as well as in history and geography for science major students. The remaining subjects are tested in the *baccalauréat* exam at the end of Grade 12. Only students who have earned the *baccalauréat* are eligible for higher education.

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

In any case, in French high schools, beyond the separation induced by the choice of an elective, gender segregation is limited. The composition of each class is determined by the principals, and, while students' electives are obviously taken into account, the principals also declare that gender is one of their top priorities (Cnesco, 2015). Most further say that they value some degree of heterogeneity in academic achievement levels, but in this area, unlike that of gender, stratification remains substantial. [2]

2.2 University Application and Enrollment

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

The higher education programs to which students can apply fall into two broad categories:

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

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

Until 2017, the admission procedure for most undergraduate programs was centralized through the *Admissions Post-Bac* (APB) online platform.[3] The main step in the procedure consisted in a variant of the college-proposing deferred acceptance mechanism (Gale and Shapley, 1962; Roth, 1982). Students were asked to submit a rank-order list of programs (ROL) that could include up to 36 choices, with a maximum of 12 choices per type of program (University program, STS, CPGE, etc.). After the submission deadline at the end of May, students were ranked by the different programs. For selective programs, the ranking was based on their students' academic records in Grade 11 and Grade 12: students' grades in the different subjects as well as teachers' written feedback were crucial role in the rankings of applicants. For non-selective programs, students were ranked according to a set of priority rules, based on their catchment area and the program's place in the student's list; that is, ranking was not based on students' grades.

3 Data and Summary Statistics

This section details the various data sources used to build our measure of gendered teacher vocabulary (GTV) and to measure the effect of exposure to a teacher with higher GTV (Section 3.1). We also present summary statistics on the sample of Grade 12 science majors and their math teachers (Section 3.2).

3.1 Data Sources

We use three main administrative databases: the higher education 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*).

[3] In 2018, a major reform of the application procedure allowed universities to select students based on their past academic performance. Since our sample period is 2012-2017, the students considered here were not affected by this new system.

APB data. Our primary source is the comprehensive application data from the APB platform over the period 2012–2017. The platform collects a substantial amount of information in the application process. First, we use the students' digitized academic records to retrieve teachers' written feedback on all the subjects taken in Grade 12 (two trimesters). This is the main input for our measure of gendered teacher vocabulary. These transcripts also report the students' grades in the continuous assessment in both Grades 11 and 12. Teachers and students are uniquely identified, so the transcripts can be matched with the characteristics of students and teachers given in a separate APB file. Along with basic sociodemographic information (gender, place and date of birth, parents' socio-economic status, etc.), the APB data provide detailed information on high school careers (school track, major and elective choices), as well as information on the teachers' gender, subject taught, and head teacher status.

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

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

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

Sample restrictions. Focusing on the feedback from math teachers, we restrict our sample to Grade 12 science majors, who naturally interact more frequently with their math teachers

than humanities or social science majors.[4] These students are also the most likely to opt for a science major at university and may therefore be more responsive to their math teacher's feedback. We exclude students for whom the math teacher's identifier or the grade transcript is missing (50 percent of Grade 12 students in the science track in 2012, but diminishing to 15 percent in 2017; Table 1). In the vast majority of cases (between 70 and 95 percent of the missing observations), the teachers' identifiers and grade transcripts are missing because the high school as such was not reporting students' grades automatically on the APB platform. Dropping these observations, that is, is tantamount to dropping entire schools and accordingly does not threaten the internal validity of our analysis.[5] Finally, we restrict our sample to high schools with at least two science major classes, since our identification strategy relies on a within-school comparison of students (see Section 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, the sample includes 40 to 75 percent of Grade 12 science major students, for a total of approximately 700,000 observations over the entire period.

[Insert Table 1 about here]

3.2 Summary Statistics

Students. Table 2 reports summary statistics of Grade 12 science majors' characteristics for the whole sample, separately for boys and girls. Students average 18 years of age and mostly come from high or a medium-high socio-economic background (43 and 16 percent respectively).[6] Girls are slightly underrepresented in the science major, making up 47 percent of science majors against 54 percent of all general academic track Grade 12 students (MENJS-MESRI, 2018). Turning to the elective courses, the gender differences are striking. Half of the girls opt for the earth and life science elective compared to only a quarter of the boys. Female students are also underrepresented in the math electives (19 percent vs. 27 percent) and engineering and computer sciences electives (6 percent vs. 20 percent). Another noticeable difference is in past academic

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

[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 all of a high school's grade transcripts are missing, regressed on the school's average characteristics. High schools with higher shares 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 is related only marginally to the probability of not reporting grade transcripts.

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

performance, as measured by the national percentile rank on the DNB exam. Boys' average rank in math is some four points higher than girls'. On the other hand, females outperform males in French at both the DNB and *baccalauréat* exams by an average of 10 percentiles. Both imbalances, in elective choices and past performance, are taken into account in our identification strategy and estimation procedure.

[Insert Table 2 about here]

Math teachers. Table 3 reports some descriptive statistics for the sample of math teachers in Grade 12 Science major courses. There are 6,772 in the sample, of whom 58 percent are men. A little more than half served as head teacher of a class at least once during our sample period. Head teachers are likely to have a stronger influence on students' performance and enrollment behavior, in that they not only teach but also counsel their students. Each teacher is in charge of only one Grade 12 science major class on average each year, with an average class-size of 28 (90 percent of teachers teach only one Grade 12 science major class per year). Teachers appear nearly four times each in our 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 teacher's feedback notes is 12.5 words, but with very considerable variability from teacher to teacher.

[Insert Table 3 about here]

4 Empirical Strategy

The first part of this section (Section 4.1) describes the estimation for measuring gendered teacher vocabulary (GTV). The second part (Section 4.2) presents the identification strategy to estimate the effect on students' outcomes of exposure to a teacher with higher GTV.

4.1 Measuring Gendered Teacher Vocabulary (GTV)

The measure of teacher GTV that we propose here leverages the rich data on teachers' written feedback to students in their Grade 12 academic records. These notes reflect the teacher's perception of students' performance, work, and behavior in class throughout the year. They are both highly relevant and highly informative to students: feedback is provided three times a year to students, is shared with their parents, and is considered by selective higher education programs during the application process. Therefore, the way in which the feedback is framed

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

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

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

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

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

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

The model described by Equation (1) is trained on a subsample of Grade 12 students. Using the set of $\hat{\alpha}$ coefficients retrieved from the estimation procedure, we use the hold-out sample to predict each student's gender as follows:[7]

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

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

In practice, the model's predictive quality, or accuracy – the proportion of correctly classified observations – could be affected by two factors that we seek to neutralize before any estimation or prediction. First, since gender is correlated with math performance (see Section 3.1), Model (1) is likely to perform better on classes with stronger gender imbalances in math. To allay this concern, for each teacher we undersample, taking the same number of boys and girls from each quartile of prior math ability (proxied by DNB percentile rank). This ensures balanced training and hold-out samples, consisting of 50 percent male and 50 percent female students from each ability level. Second, feedback length varies substantially among teachers (see Table 3): obviously, a longer note, with more words used, is mechanically likely to generate more accurate predictions. For feedback of above-median length, we circumvent this issue by randomly sampling 12 words (the median number).

Estimating Model (1) on the balanced subsample yields an accuracy of 63 percent: the model predicts the student's gender correctly in 63 percent of the cases. It performs slightly better at predicting male than female gender (65 vs. 63 percent).[8]

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

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

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

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

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

Both measures range between 0 (the model systematically misclassifies females as males

[8]We tried more flexible specifications of the model by supplementing single words (unigrams) with interactions between words (bigrams), but this more complex specification did not improve predictive quality.

and males as females) and 100 (all students are assigned their actual gender). The greater the accuracy for a given teacher, the better we can recover their students' gender based on the words used in the feedback, and hence the stronger the gender differentiation in the vocabulary used in these assessments. A model that assigned student gender randomly with probability 0.5, would achieve a 50 percent accuracy. Thus, our model predicts gender better than random guessing for all teachers whose accuracy is above 50 percent[9].

4.2 Identification Strategy

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

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

where Y_{isjet} is the outcome of student i in high school s with elective courses e taught by teacher j during academic year t. $GTV_{j\backslash c}$ is teacher j's standardized GTV measure and is class-specific, in that we use the *leave-one-out* GTV described in Equation (5). The coefficient γ_{set} is a set of school × elective course × year fixed effects. Hereafter GTV is standardized, and the coefficient of interest is β_1, which measures how a student's outcome is affected by a math teacher with a 1-standard-deviation higher GTV. The standard errors are robust and clustered at teacher level.[10] For this identification strategy to be valid, the *leave-one-out* GTV must not be systematically correlated with students' characteristics. We test this formally in Section 6.

5 Gendered Math Feedback

We first report the distribution of our GTV measures (Section 5.1) and then present a qualitative analysis of the gendered vocabulary used (Section 5.2).

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

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

5.1 The distribution of the GTV measures

Figure 1 shows the density and the cumulative distributions of the GTV and the *leave-one-out* GTV measures separately. There is evidence of a correlation between students' gender and the math feedback received, controlling for previously demonstrated math aptitude. Our model predicts gender better than random guessing for 90 percent of math teachers using the standard GTV measure and for over 95 percent using the *leave-one-out* GTV.[11] For the median teacher, the student's gender is predicted correctly in 63 percent of the time. By comparison, the model achieves median accuracy of 66 percent in predicting the student's math performance, which is the upper bound of what we could expect given that the feedback serves precisely to assess this performance.[12] Breaking the GTV distributions down by teacher's gender, women math teachers differentiate their vocabulary slightly more, on average, than their male colleagues (see Figure C3).

[Insert Figure 1 about here]

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

[Insert Figure 2 about here]

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

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

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

[14] The median leave-one-out GTVs 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.

5.2 Qualitative Analysis of the Best Gender Predictors

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

To gauge the extent to which the gendered vocabulary expresses gendered beliefs and expectations, we explore the actual feedback content analyzing the best gender predictors. Building on the psychology literature on the classification of teacher feedback and mindsets (Morgan, 2001; Burnett, 2002; Dweck, 2006), we classify the gender predictors into five different categories to capture the different beliefs and expectations conveyed. First, depending on the valence of the word, we categorize it as positive, neutral or negative. Second, words referring to students' attitude in class or the effort dedicated to the subject are classified as "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 either category remain unclassified.[15] The classification of the top 100 male and female predictors is shown in Appendix Tables B2 and B3.[16]

Analysis of the best gender predictors. The analysis reveals marked differences in the qualifiers used by teachers according to the student's gender. Figure 3 reports the odds ratios derived from the estimation of the model described by Equation (1) for the top 10 predictors of each gender. Feedback referring to lack of confidence, propensity to get discouraged or cheerful aspect ("smiling") is between 1.8 and 2.1 times more likely to be directed to a girl than to a boy, relative to other feedback. Teachers are also more likely to note that female students are stressed or panicked, and to cite their exemplary conduct ("exemplary", "studious"). On the other hand, feedback that describes the student as childish ("childish", "has fun"), comments

[15] We sought to confirm our classification with data-driven techniques using bi-term topic models tailored for short texts. However, these models performed poorly on our data, which are highly specific, in that the texts are very short, averaging just 12 tokens per teacher, and the overall vocabulary is quite limited (on average 1,600 words), with little variation in the topics (almost all relating to academic performance and behavior). This presented the typical challenges inherent in such short texts: the topics generated gathered inconsistent words (*trivial topics*) and the different topics were highly similar with a large share of words in common (*repetitive topics*, see Wu et al. 2020 for a discussion of these issues.)

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

on the need for careful handwriting, or praises the student's curiosity and intuitions is between 1.8 and 2.3 times more likely to be addressed to by a male than by a female student, compared to feedback that does not mention these terms.

[Insert Figure 3 about here]

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 feedback by gender. Among the top 30 male predictors, only 8 are positive, while two thirds of the best female predictors can be classed as positive. Most interestingly, conditional on being positive, almost all the best male predictors are competence-related ("curious", " idea", "interest", "intuition"), while nearly all the best female predictors refer to managerial aspects ("irreproachable", "willingness", "persistent"). On the other hand, over 75 percent of the best male predictors can be classified as negative, referring overwhelmingly to disruptive behavior ("has fun", "childish") or to lack of work-effort ("waste", "superficial").[17]

[Insert Figure 4 about here]

Analyzing all the gender predictors together, the foregoing results stand confirmed. Figure 5 shows the proportions of negative, positive and neutral feedback, conditional on whether it is competence-related (Panel A) or behavior-related (Panel B). Panel A shows that only 20 percent of the female predictors that can be classified as competence-related are positive, compared with 38 percent for male predictors, while 17 percent are negative (11 percent for male students), and the rest are neutral. Symmetrically, among the female predictors classified as managerial, 44 percent correspond to positive feedback compared with 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, as against just only 31 percent for females.

Turning next to the proportions of competence, managerial and neutral words used in positive and negative feedback (Panel B), we find that conditional on being positive, the top female predictors qualify their competence-related skills in only 17 percent of cases compared with 39 percent for their male counterparts. For negative predictors, 38 percent of the male predictors and 36 percent of the female predictors are managerial, for negative competence-related predictors these proportions are respectively 16 percent and 9 percent.

[17]The classification of the top 30 gender predictors when bigrams are used instead of unigrams is displayed in Appendix Figure B1; the conclusions are not altered.

[Insert Figure 5 about here]

Vocabulary according to GTV decile. We further investigate whether teachers with varying degrees of GTV differ, by comparing 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 teacher gender gaps and Panel (b) displays the share with a gender gap in favor of girls, separately for competence- and managerial-related feedback. The gender gaps in the share of positive words rise at an increasing pace with the GTV decile, indicating that teachers with higher GTV tend to provide relatively more positive feedback to one gender over the other, and the more so, the higher the GTV decile. This is true for both managerial-related and competence-related feedback, with a gender gap that widens from 6 or 7 percentage points in the lower GTV deciles to 10 points in the top decile. In line with the findings from the analysis of the gender predictors, Panel (b) shows that relatively high-GTV teachers are overwhelmingly more positive towards female students in their managerial-related feedback, more negative in their competence-related feedback.

[Insert Figure 6 about here]

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

6 The Impact of Gendered Feedback on Student Outcomes

Having found significant differences in the gendered vocabulary of Grade 12 math teachers, we now turn to their effect on students' outcomes. First, we report a series of statistical tests to validate our empirical strategy (Section 6.1). We then discuss how the exposure to teachers with different levels of GTV affects academic performance, higher education application behavior and enrollment in the year after high school graduation (Section 6.2). Our results prove to be robust to a series of alternative specifications.

6.1 The Validity of the Empirical Strategy

6.1.1 Exogeneity Assumption

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

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

[Insert Table 4 about here]

Random allocation of students. To check whether students are allocated randomly to teachers within a given high school, elective course and year, we perform a series of Pearson's Chi-square tests of independence. For each unique combination of school, elective course, and year, we tabulate math teachers' identifiers with each of the students' baseline characteristics and test for independence.[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 the nominal levels). For the female dummy, the empirical p-value is 11 percent. That is, in 11 percent of every 100 high school×elective×year combinations, we cannot exclude the non-random assignment of female students to classes at the 95 percent level. To ensure that the results in Section 6.2 are not driven by these slight gender imbalances, Equation (6) is also estimated with the average

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

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

proportion of females in the class as an additional control, as well as with the full set of students' baseline characteristics.

[Insert Table 5 about here]

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

6.1.2 Reverse Causality

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

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

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

Third, looking at the distributions of *leave-one-out* GTV measures estimated for other subjects further highlights the specific nature of science subjects, for which gender is better predicted on average than for the humanities (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, suggesting that, for a given class, science teachers tend to differentiate their vocabulary more. This means that our measure captures differences that go beyond characteristics specific to a class.

Finally, the fact that teachers' GTV is computed for multiple years and classes makes it possible to measure persistence of GTV across classes and over time. The correlation between a teacher's GTV and the *leave-one-out* GTV, i.e., the correlation between a given GTV and its average computed in other years×classes, is 0.161 and is significantly different from zero.[20] Note, however, that this correlation suffers from an attenuation bias, because we are correlating several GTVs measured with error, owing to the small sample used for the prediction at the class level. By way of comparison, in the teacher value-added literature, the within-teacher

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

correlation is usually around 0.3 (Chetty et al., 2014). All in all, we are convinced that our GTV measure captures persistent differences between the GTV of different teachers.

6.2 The Effect on Student Performance and Enrollment

Academic Performance. Panel A of Table 6 reports the estimated effect of having a teacher with a higher *leave-one-out* GTV on students' standardized math performance on the *baccalauréat*, based on Equation (6). On average, being exposed to a teacher with a 1-standard-deviation higher GTV improves math performance on the *baccalauréat* by 1.6 percent of a standard deviation on average, a value significant at the 1 percent level. This effect corresponds to moving from a teacher with an average GTV to one at the 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 than for male students (1.4 percent). 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 in physics, biology and philosophy. The effects are not statistically significant for any of these core subjects, which is consistent with the idea that our baseline estimates capture the effect of the math teacher and not some unobserved differences between classes.

[Insert Table 6 about here]

These moderate average effects conceal a heterogeneity of responses depending on the degree of gender differentiation in the math teacher's feedback notes. Rather than include teacher GTV in the equation linearly, we explore the intensity of the treatment by regressing students' math grade on a set of GTV deciles, the first decile corresponding to the bottom 10 percent of the leave-one-out GTV distribution. Figure 7 plots the coefficients associated with the GTV deciles along with their 95 percent confidence intervals, separately for boys and girls. Compared with students exposed to the bottom 10 percent of teachers in terms of GTV, those with teachers in the 4^{th} decile or above perform better on the *baccalauréat* exam 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 according to students' previous math performance or socio-economic status (see Appendix E for details).

[Insert Figure 7 about here]

University applications and enrollment. Although having a math teacher with 1-standard-deviation higher GTV significantly improves female and male students' performance,

we find no evidence of significant effects on their university application and enrollment outcomes. The effects on the probability of students' taking a STEM program as top choice in their applications (Table 6, Panel B) and of enrollment in a STEM undergraduate program (Table 6, Panel C) are small and statistically insignificant at conventional levels. If anything, male students are marginally less likely (-0.4 percentage point) to top rank and enroll in a selective STEM program, a 1.9- percent decrease from the baseline probability of 21 percent. We find no evidence of any heterogeneous effects by decile of teacher GTV, by previous math performance or by socio-economic status.

Robustness checks. Our results are robust to a series of sensitivity tests (data 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 (see Section 6.1), we estimate the model in Equation (6) controlling for students' baseline characteristics (columns 1 and 4) and for the share of girls in the class (columns 2 and 5). Second, to strengthen the argument that the effect we estimate is specific to math teacher GTV, we control for the average GTV in the same class for the five other core subjects (Columns 3 and 6). Including these controls does not alter the magnitude or the significance of the results. The effect on math performance of a 1-standard-deviation increase in GTV ranges from 1.2 to 1.4 percent of a standard deviation for boys, and 1.8 to 2.1 percent for girls. The limited but significant effects on the probability of top-ranking a STEM program in the ROL and on the probability of enrolling in a STEM program (a decrease of about 0.5 percentage point) persist among boys.

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 could be correlated with other teacher characteristics. The estimated coefficients, that is, might capture the effects of the latter. Here, we explore the possible mechanisms that could drive the effects of teachers' GTV on math performance, showing that the effects of gendered feedback remain even after accounting for the potential confounders. First, we investigate whether teachers using a gendered vocabulary also encourage girls by overgrading them relative to boys (Section 7.1). Second, we investigate whether teachers using gendered feedback are also more likely to personalize their feedback notes (Section 7.2). Third, we compute a measure of teacher quality to see whether math teachers with a higher GTV are also better teachers (section 7.3). Finally, we test whether

the performance effect is different, for a given level of gendered feedback, when the teacher is more likely to use either male-specific or female-specific vocabulary (Section 7.4).[21]

7.1 Teacher Grading Bias

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

As to the correlation between grading bias and our measure of GTV, a 1-standard-deviation increase in GTV is associated with a 0.06-standard deviation decrease in the grading bias, significant at the 1 percent level (Panel (a) of Figure F8). This correlation is low but nevertheless suggests that teachers who use a more highly gendered vocabulary may also be slightly more likely to encourage female students with higher continuous assessment grades. However, 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 estimated coefficients on GTV and on teachers' grading bias. A 1-standard-deviation increase in grading bias increases math performance by 1.4 percent of a standard deviation for boys and 1.2 percent for girls, a magnitude comparable to our effects of exposure to gendered feedback. But, its inclusion as a control does not alter the estimated coefficients on GTV, so we can rule out grading bias as a first-order mediator of the impact of GTV.

[Insert Table 7 about here]

7.2 Teacher Feedback Personalization

Next, we investigate whether gendered feedback has an effect on student outcomes beyond that of simple feedback personalization as such, which has been shown to reinforce motivation and to enhance students' performance (Koenka and Anderman, 2019). Feedback personalization is likely

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

to be related to the GTV measure, as less feedback personalization may mechanically decrease the probability that the model will detect gendered feedback. To take an extreme example, a teacher who copy-pastes his notes for students within some given 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 might be capturing the effects of feedback personalization.

To investigate this issue, we construct 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 notes provided to each pair of students.[22] The greater the text distance between the different feedback notes, the more the teacher personalizes feedback. Figure F6 displays the distribution of teacher text distance and its correlation with leave-one-out GTV. While the correlation is 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 estimating the model in Equation (6) does not affect the relationship between GTV and student performance. While exposure to a teacher with personalized feedback appears to be beneficial to both boys and girls, having a teacher whose feedback is gendered still improves math performance significantly, by 1.4 percent of a standard deviation for boys and 2.1 percent for girls (Panel b of Table 7). Therefore, we can discard the thesis that having a teacher with higher GTV affects performance through feedback personalization rather than through gender differentiation.

7.3 Teacher Quality

We next investigate whether benefits of exposure to a higher-GTV teacher might not be mediated by differences in teacher quality as such. 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 baseline student characteristics, measures of previous 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 resulting coefficients 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 are given in Appendix F.

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

We then check whether GTV is correlated with teachers' quality. Panel (b) of Appendix Figure F8 suggests a small but significant quadratic relationship, where teachers with GTV measures 2 standard deviations below or above the average have slightly lower value-added.

Controlling for teacher quality in Equation (6), we find that the effect of gendered feedback on math performance is reduced by 0.6 percentage points for boys and 0.7 for girls compared with the specification without this control: from +1.4 to +0.8 percent of a standard deviation in math grade for boys and from +2.1 to +1.4 percent for girls (Table 7, Panel c). That is, the effect of higher GTV on students' math performance, while reduced, remains significant, suggesting that our results depend only in part on the teacher quality mechanism.

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

The last mechanism we explore is whether the type of vocabulary used by the teacher, i.e., the words more likely to be used for girls or boys, triggers different responses from students. While the general female-specific and male-specific vocabulary can be described and classified in broad categories, as in Section 5, this cannot be done at the teacher level, so as to properly distinguish different feedback styles (e.g. "encouraging", "competence-related"). We seek to disentangle the effect of exposure to a teacher who is relatively more likely to use the "female vocabulary" (or the "male vocabulary") for a given level of gendered feedback.

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

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. Male students are more often correctly classified by our model (65 percent 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 correctly predicted frequently have a substantially lower proportion of correct classifications for the other gender. A 1-standard-deviation increase in GTV-male is associated with a 0.7-standard-deviation decrease in GTV-female. This suggests that teachers using the female-specific vocabulary for their female students are not systematically

using the male-specific vocabulary for males, but female or gender-neutral vocabulary.

[Insert Table 8 about here]

To measure whether the positive effect on math performance at *baccalauréat* is reinforced by teachers more likely to use the vocabulary associated with female or with male students, we estimate Equation (6) augmented with GTV-male (or GTV-female) in addition to the overall GTV measure (Table 8). The positive effect on math performance documented for higher-GTV teachers is reinforced for teachers with a higher GTV-female (those more likely to use the vocabulary associated with females). For a given level of GTV, a 1-standard-deviation increase in GTV-female is associated with an average additional increase in math performance of 0.6 percent of a standard deviation. Our results suggest that exposure to a higher GTV-female teacher matters only for girls, for whom the standardized math grade increases by an additional 0.9 percent of a standard deviation against a non-significant 0.3 percent for boys. On the other hand, exposure to higher GTV-male teachers diminishes the positive effect on math performance and seems to be detrimental to female students, whose math performance worsens by 0.8 percent of a standard deviation, while male students are largely unaffected.

Altogether, investigation of the various mechanisms hypothesized here suggests that exposure to gendered feedback affects students' performance, over and above other teacher characteristics. These findings 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 the student's gender affects the vocabulary used by math teachers in their written feedback. To identify gendered patterns in teacher feedback notes, we apply machine learning techniques to predict students' gender from the words that the teachers elect. The key findings are threefold. First, we find that equally able female and male students get different feedback, and that the scope of this gender differentiation is ample: overall, the words used by Grade 12 math teachers allow correct prediction of the sex of 63 percent of the students. For comparison, this is only marginally lower than the feedback's predictive power as regards student performance, which is the upper bound of what we could expect given that the express purpose of the feedback is to assess performance. Second, qualitative analysis of the best gender predictors reveals that for female students, teachers tend to emphasize positive managerial

aspects and to encourage effort, while equally performing males are both more severely criticized for unruly behavior and more fulsomely praised for intellectual skills.

We take the analysis one step further by investigating how gendered feedback relates to student performance, college applications and enrollment decisions in the following year. We compute a teacher-level measure of gendered teacher vocabulary (GTV), which we define as the share of students whose gender is correctly predicted. Exploiting the quasi-random assignment of teachers within high schools conditional on the elective courses chosen by students, we relate GTV to educational outcomes. The third key finding is that exposure to a teacher with higher GTV improves math performance by between 1.6 percent and 6 percent of a standard deviation, with slightly larger effects for girls than for boys. We find no significant or sizeable effects on higher education applications or enrollment decisions in the subsequent year. Given the rather limited effects on math performance, this absence of effects on university outcomes should not be surprising, since performance itself is the main mediator between GTV and these outcomes. However, one might have expected GTV to influence university outcomes by modifying aspirations and self-confidence. The lack of any such effect suggests that this is not the case.

The magnitude of the effect of exposure to gendered feedback is within the lower bound of what this and other work has found for other input factors in the education production function. As to the effect of teacher quality on test scores –which constitutes an upper bound to the expected effect on student performance– Chetty et al. (2014) find that a 1-standard-deviation increase in teacher value-added improves math test scores by 14 percent of a standard deviation, while in our setting estimates range between 16 percent and 20 percent. In the grading bias literature, Terrier (2020), for instance, finds 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 1-standard-deviation stronger implicit stereotypes widens the gender gap in math by 4 percent of a standard deviation, more comparable to the findings of the present paper.

Finally, we explore a range of potential mechanisms for the effects on math performance. We provide suggestive evidence that gendered feedback as such is important, by controlling for three teacher characteristics that are likely to be correlated with GTV: gender grading bias, personalized feedback, and teacher quality. Second, seeking to get inside the black box of gendered feedback to understand which aspects of feedback potentially may help to determine student performance, we compare the effects of exposure to teachers who make greater or lesser use of the vocabulary associated with female or male students, for a given level of gendered

feedback. We find that the effect of gendered feedback on boys' math performance does not vary with the type of vocabulary, but for girls it is greater when the teacher uses the vocabulary associated with female students, i.e. when teachers underscore the girls' positive behavior and effort.

The main take-away from our study should be an awareness message. Our findings indicate that teachers' written feedback may be an effective pedagogical device to improve students' performance. The paper implicitly suggests avenues for future research. Since our setting does not involve feedback manipulation, but only manipulation in exposure to teachers with different levels of gendered feedback, we cannot properly identify the features of gendered vocabulary that trigger the strongest impact on performance. To go a step further, we need additional research using experimental settings in which teacher feedback types vary randomly.

References

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.*

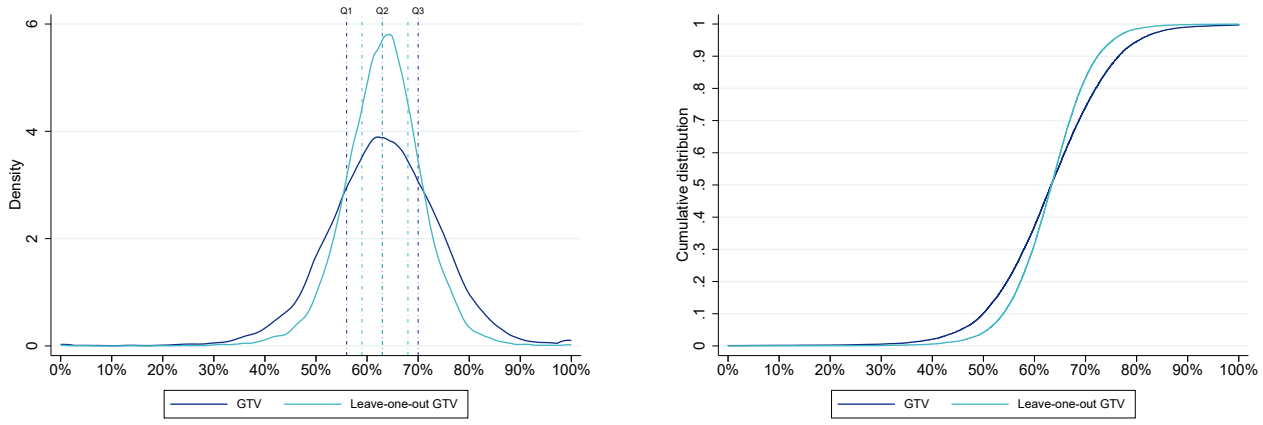
Terrier, C., "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement," *Economics of Education Review*, 2020, *77.*

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.

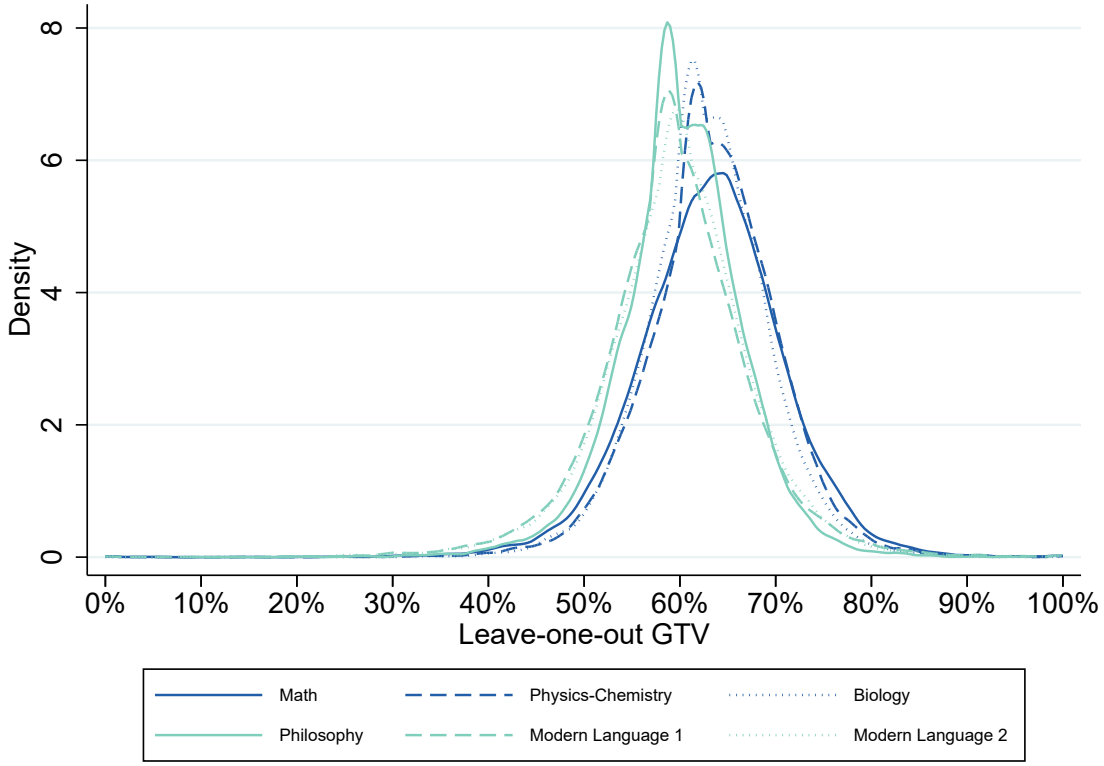
Figure 1 – Distribution of Math Teachers GTV and Leave-one-out GTV



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

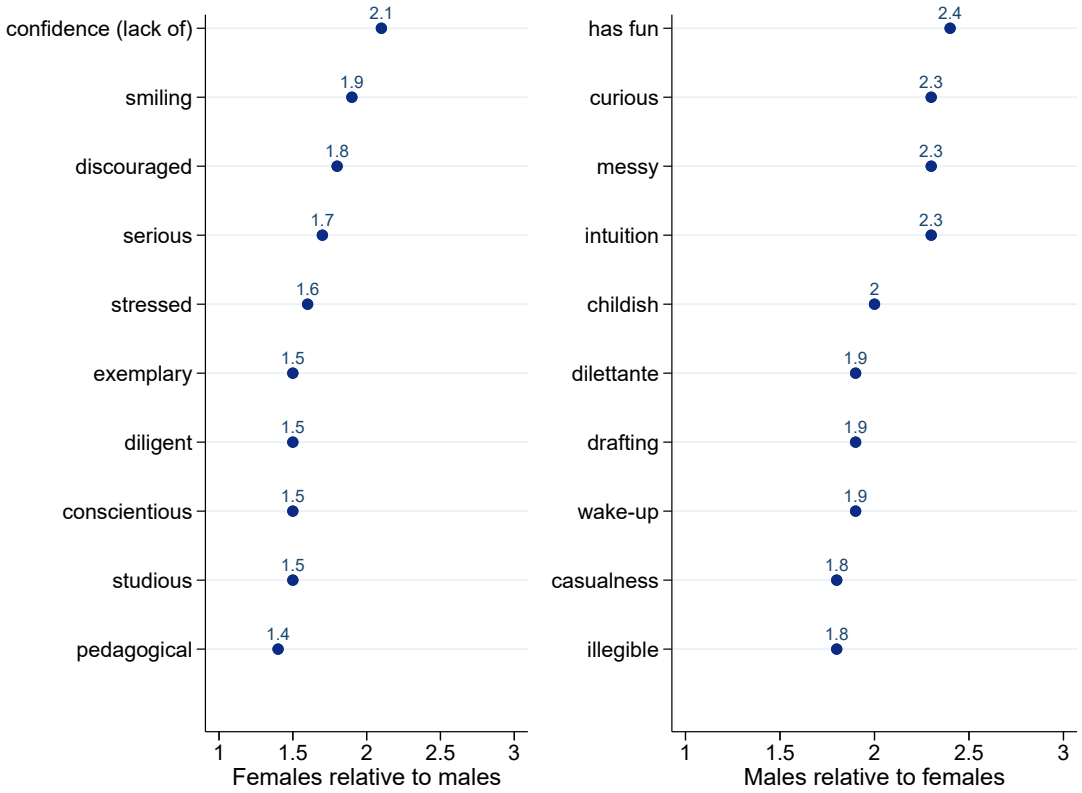
Notes: This figure shows the densities (Panel (a)) and cumulative distributions (Panel (b)) of the math teachers' GTV and leave-one-out GTV measures. The vertical lines in Panel (a) represent the first, second and third quartiles of the GTV distributions. Computations are based on administrative data 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



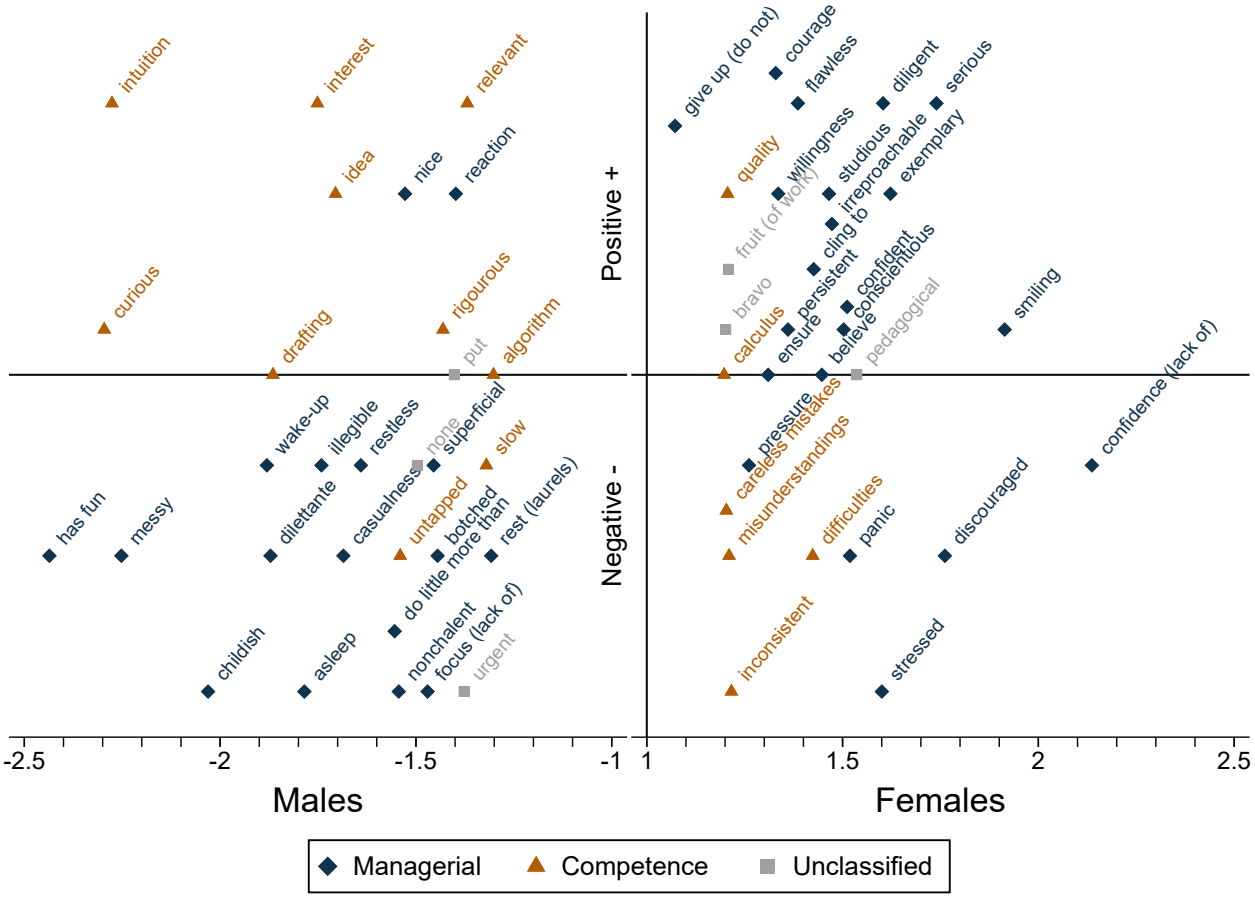
Notes: This figure shows the distributions of the math, physics, biology, philosophy and foreign language teachers' leave-one-out GTV measure, based on administrative data from the French Ministry of Education. The sample consists of Grade 12 teachers teaching in high school × elective × year cells containing 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



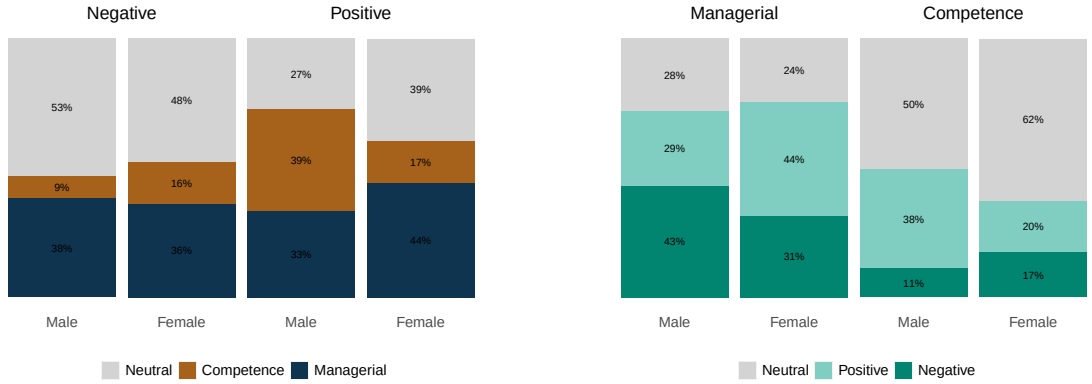
Notes: This figure shows the odds ratios obtained for the 10 best predictors of the female gender (left-hand side), and for the 10 best predictors of the male gender (right-hand side). These odds ratios are obtained from the estimation of the model described by Equation (1), where the vocabulary appearing 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.

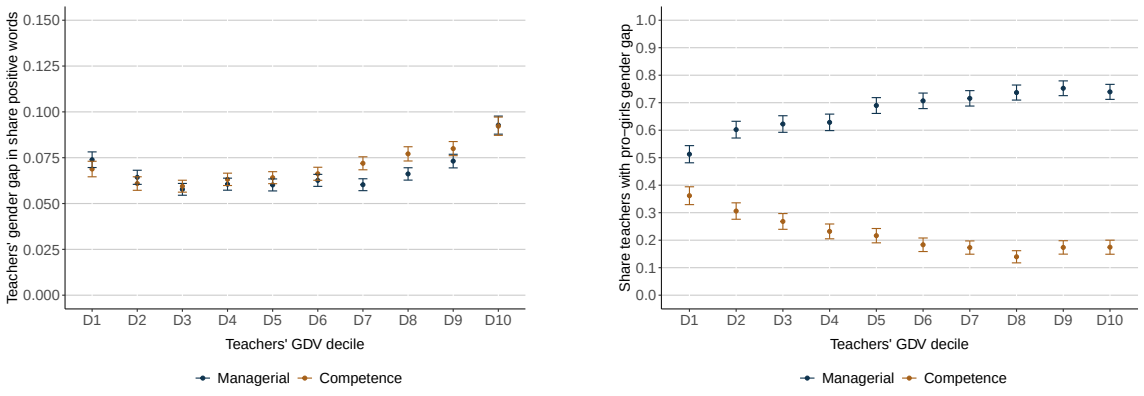
Figure 5 – Gender Predictors' Type and Positiveness



(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.

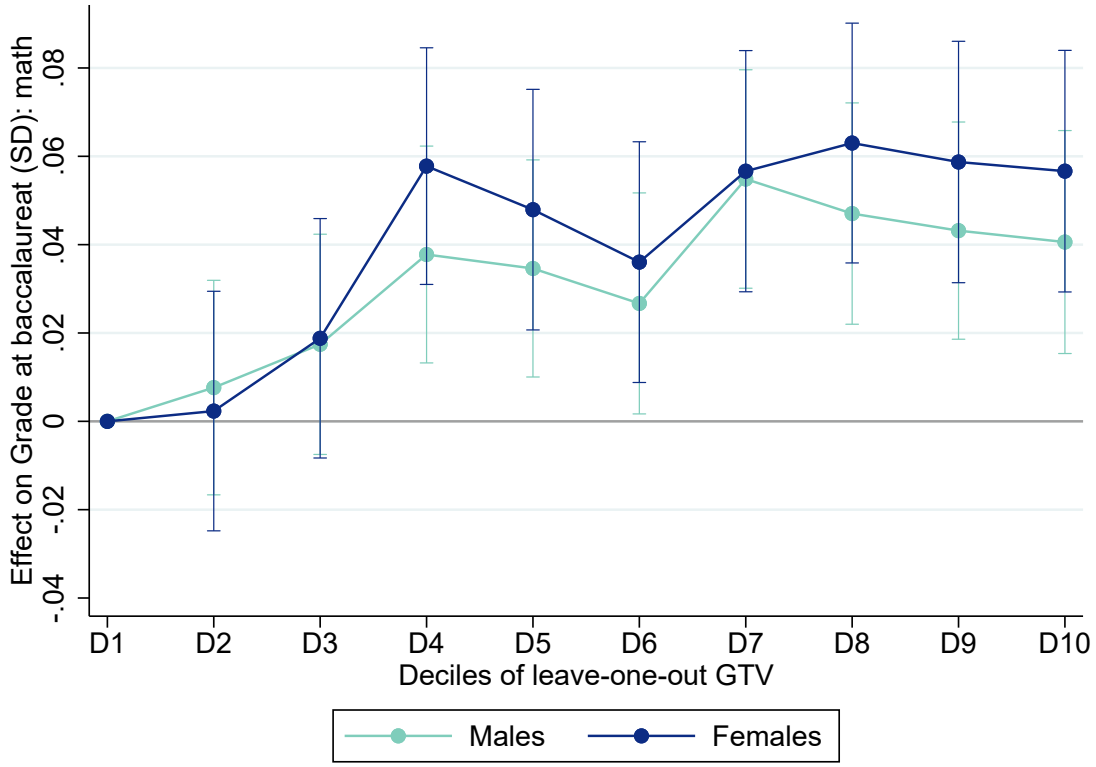
Figure 6 – Teachers' Gender Gap in the Share of Positive Words in Favor of Females by Deciles of GTV - In Absolute Value and Percentage



(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	**Boys**	**Girls**
	(1)	(2)	(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

October 2022

List of Appendices

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

This appendix details the practical implementation of the steps taken in estimating the gendered teacher vocabulary (GTV) 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 note a teacher wrote to a given student, in a given subject. Our aim is to convert all the documents into a data structure similar to the one displayed in Table A1. In this example, all the words and groupings of two words that appear at least once in a document have been converted to a column.

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

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

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

Table A1 – From text to data: an illustration

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

A.2 Predicting Student Gender and Measuring GTV

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

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

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

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

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

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

Step 2: Random selection of tokens. As shown in Table 3, the number of tokens used in feedback varies by teacher. As feedback length could influence the quality of the prediction, we randomly sample twelve tokens for lengthy feedback, defined as the ones with an above-median length (12 words).

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

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

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

[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.

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:

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

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

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

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

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

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

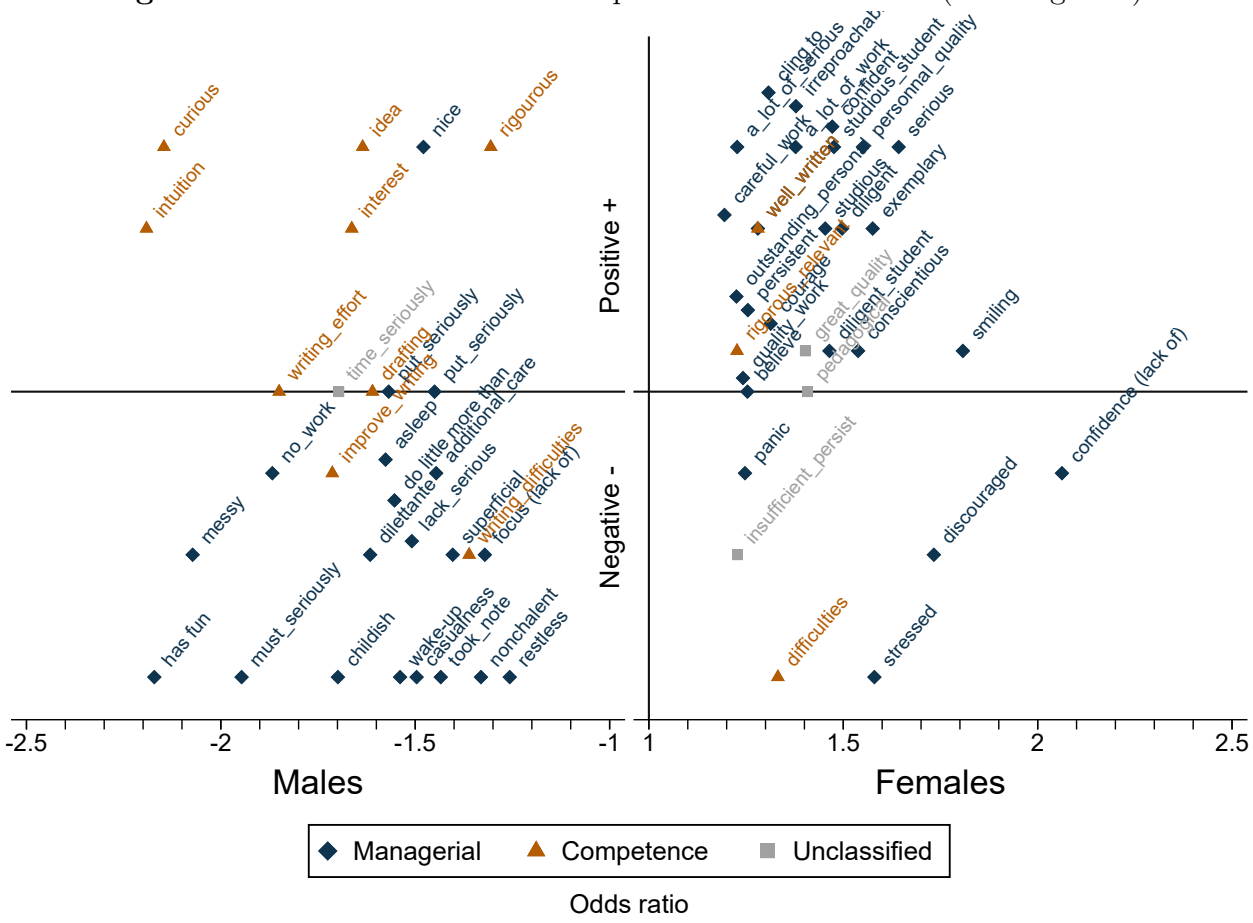
	Positive	**Negative**	**Neutral**
Competence-related	**4 tokens:** accurate, autonomous, master, quality	**6 tokens:** careless mistakes, difficulties, inconsistent, mishap, mistake, misunderstandings	**11 tokens:** appropriate, assessment, calculus, elementary, litteral, method, methodological, question, 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

ht

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.

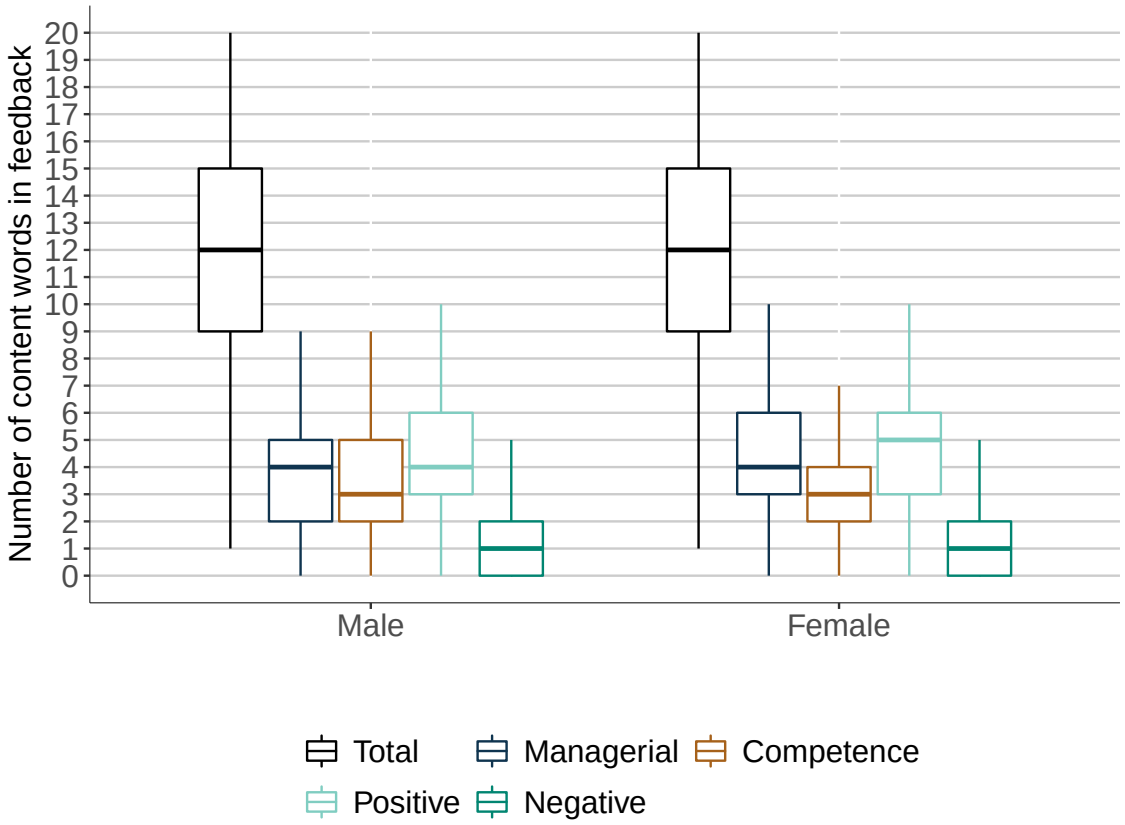
B.2 Classification of the Top 30 Gender Predictors - Bigrams

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 math feedback notes' word counts 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 notes of the same length with a median 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 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.

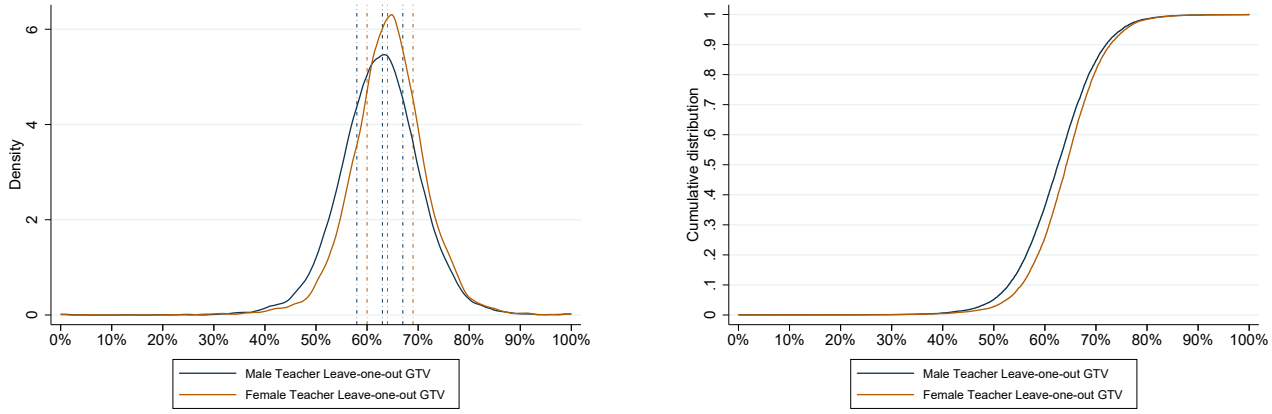
Figure B2 – Math Feedback - Distribution of Word Counts



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.

ht

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 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

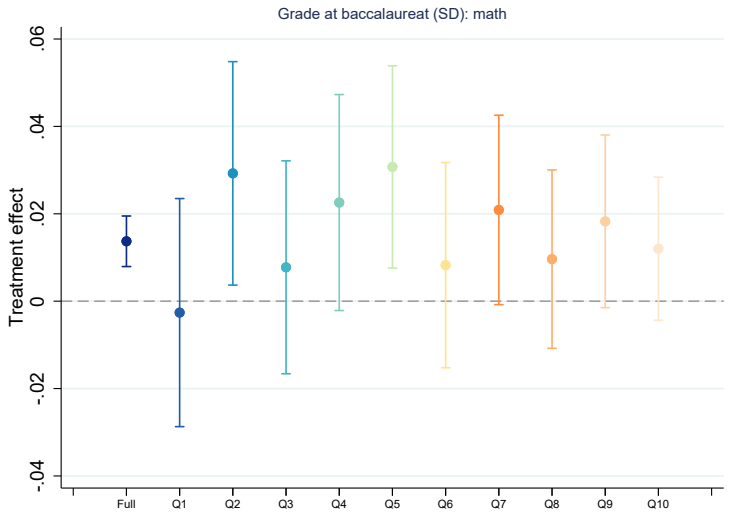
Table E7 – Effects of Exposure to Gendered Teacher Vocabulary (GTV) on Students' Educational Outcomes - Robustness Checks

	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	

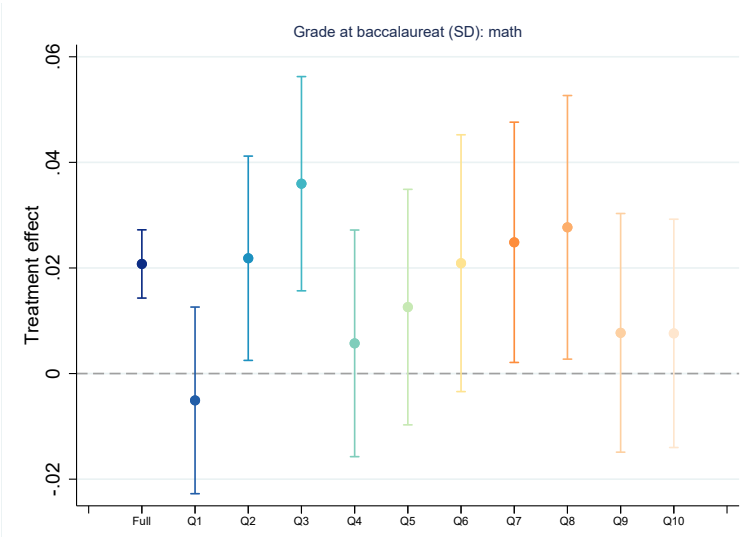
Notes: Each row reports the coefficients on the standardized *leave-one-out* teacher GTV obtained from the estimation of Equation (6) for the different outcomes listed on the first column. It is estimated on the whole sample and separately for Grade 12 male and female students. The regression includes 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) Boys

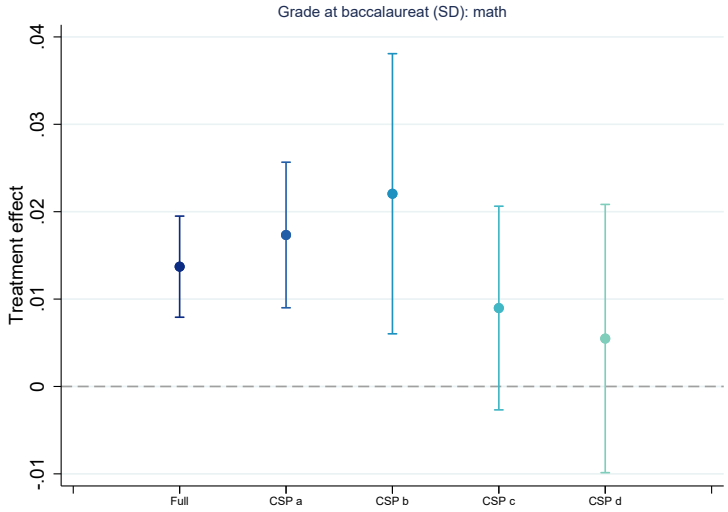


(b) Girls

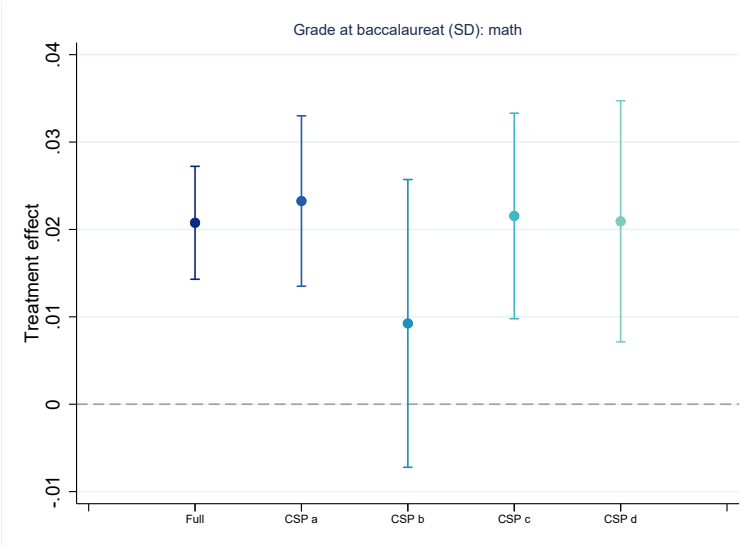
Notes: The figure reports the effect of a one standard deviation increase in leave-one-out GTV on students' standardized grade on the math *baccalauréat* exam separately by gender and by initial performance in math. Initial math performance is measured as deciles of percentile rank in math obtained at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

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) Boys



(b) Girls

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

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 on the continuous assessment as the non-blind score, and the (standardized) math grade obtained on the *baccalauréat* exam as the blind score. The grading bias (GB) for class c taught by teacher j in year t is therefore defined as follows:

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

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

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

Table F8 – Maths grades during G12 and at Baccalaureat exam - By students' and teachers' gender

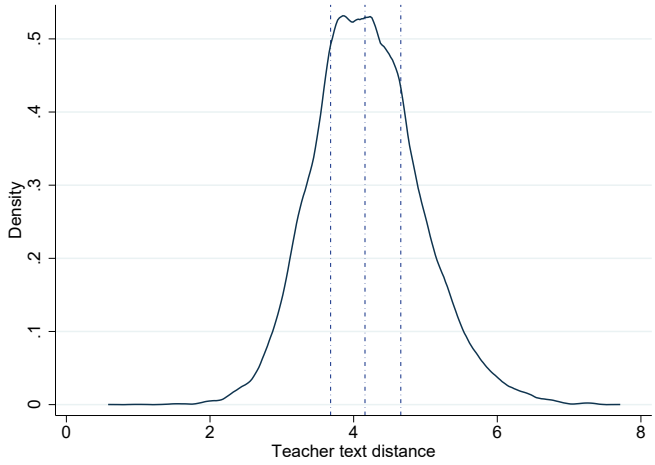
	Boys			**Girls**			**Teacher**
	G12 maths	Bac maths	Diff.	G12 maths	Bac maths	Diff.	**bias**
All teachers	−0.017	0.043	−0.061	0.020	−0.049	0.068	−0.129
Female teachers	−0.029	0.028	−0.057	0.033	−0.031	0.064	−0.121
Male teachers	−0.009	0.054	−0.063	0.010	−0.062	0.072	−0.135
N	364,769	344,131		319,552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

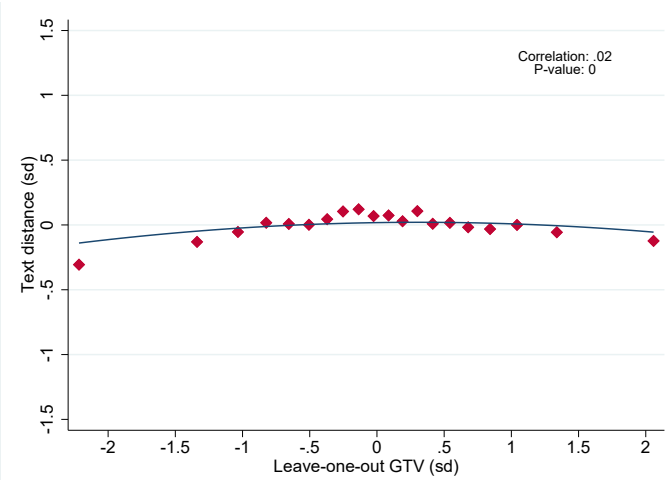
Figure F8 displays the correlation between our measure of GTV and the teacher grading bias. While statistically significant, the magnitude of the correlation between the two standardized measure is negligeable.

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 Chetty et al. (2014). 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 regress students' test scores in year t, measured by the percentile rank obtained on the math *baccalauréat* exam, on a set of students' baseline covariates, controls for students' previous 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 teacher 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.

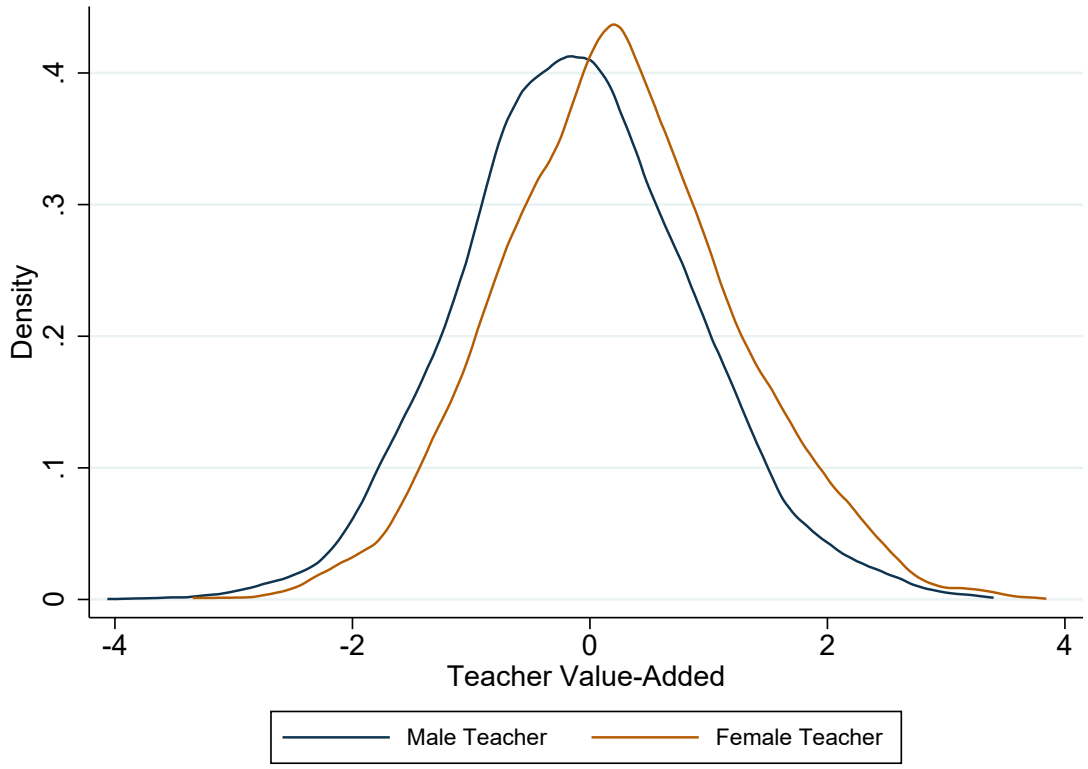
Step 3: Predicting teachers' quality. The final step consists in using the set of OLS coefficients from step 2 to *predict* teachers' quality. This predicted teacher quality is actually just a proxy for a teacher's true value-added and its reliability depends on the shrinkage factor, usually estimated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

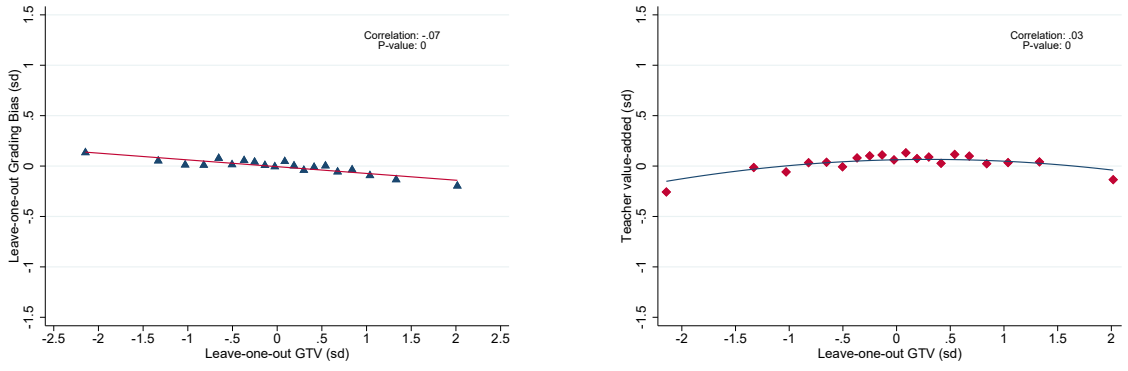
[A.2] Teacher fixed effects are included in the regression so that coefficients on other covariates are estimated only using the within teacher variation. Those fixed effects are then added back to the residuals.

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

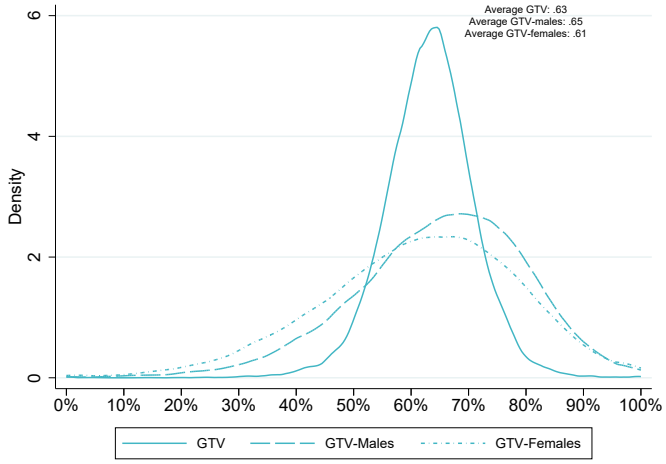


(a) Teacher GTV and Teacher Grading Bias **(b)** Teacher GTV and Teacher Value-Added

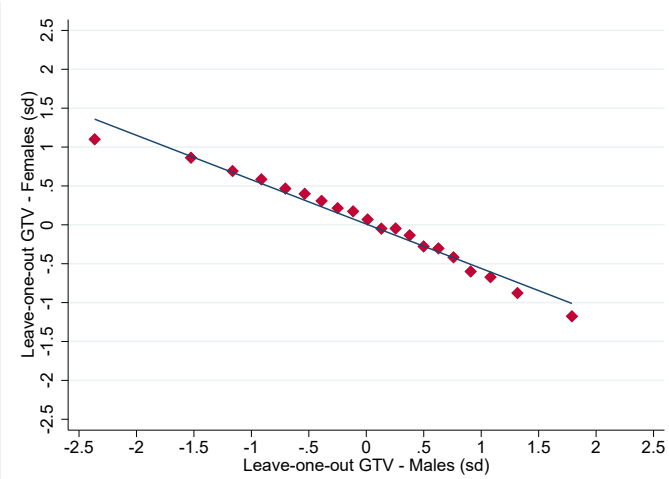
Notes: The figure shows the binned average of the teachers' leave-one-out grading bias (resp. value-added) standardised measures on the standardised teacher leave-one-out GTV. The line represents the linear fit in Panel (a) and the quadratic fit in Panel (b). The correlation coefficients are obtained 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.