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

Pauline Charousset
Marion Monnet

May 2021
(Preliminary version – Please do not circulate)

Abstract

We propose a novel approach to measure the degree of gender differentiation in the vocabulary that teachers use to assess their students' work. We build a model that predicts each student's gender based on the words appearing in the feedback she receives. For every teacher, we estimate a synthetic measure of gender-differentiation by computing the model's predictive accuracy on each teacher's students, controlling for gender imbalances in ability. Using the universe of French Grade 12 students' math transcripts, we show that math teachers use different words to assess the work of male and female students with a similar performance level. The analysis of our model's best gender predictors reveals that teachers insist more on positive managerial aspects for their female students, while equally performing males are both more criticized for their unruly behavior and praised for their intellectual skills. Using the within high school, elective and year variation, we further show that a higher gender differentiation in the teachers' feedback translates into an improved performance at the math *baccalauréat* exam, but does not affect students' matriculation in the following year.

JEL codes: I21, I24, J16.
Keywords: *higher education*, *teacher feedback*, *selective STEM*

[*] Charousset: Paris School of Economics and Paris 1, 48 boulevard Jourdan, 75014, Paris, France (e-mail: pauline.charousset@gmail.com); Monnet: French National Institute of Demographic Studies (Ined), Paris, France (e-mail: marion.monnet@ined.fr).

1 Introduction

Stereotypes about different gender aptitudes in science-related fields are still pervasive and are now widely acknowledged as one of the causes of women's underrepresentation in Science, Technology, Engineering and Math (STEM) fields (Miller et al., 2014; Leslie et al., 2015). Stereotyping STEM as a male domain influences females' attitudes towards science, their performance in science subjects (Nosek et al., 2009), and eventually their interest as well as their willingness to pursue a scientific career (Cheryan and Plaut, 2010).

The development of gender-related science beliefs starts at early ages (Bian et al., 2017), first shaped by parents' attitudes and beliefs towards science and, later on, by the schooling environment, namely teachers. Several studies show that math teachers hold stereotyped beliefs conveyed to their students through their classroom instructions (Keller, 2001), their own attitudes such as math-anxiety (Beilock et al., 2010), or through the type of feedback provided.[1] Teachers holding traditional gender views tend to adopt more person-based feedback insisting on fixed aptitudes (e.g. "You are gifted in math"). They convey the idea that math ability is innate and that females are less likely to have it (Copur-Gencturk et al., 2020), while more progressive teachers use effort-based feedback, with the idea that math ability can be acquired with efforts (e.g. "You have worked hard"). The type of feedback provided sends a signal about whether intelligence is fixed or malleable and is all the more crucial as females are more sensitive than males to the different feedback received. Effort-based feedback enhances females' motivation (Corpus and Lepper, 2007), their sense of belonging and willingness to pursue math (Good et al., 2012), while those aspects are harmed by person-based feedback.

This paper documents the extent to which the feedback used by math teachers is gender-differentiated, and how exposure to teachers using such a gender-differentiated feedback affects Grade 12 students' academic performance and higher education enrollment choices.

To our knowledge, the analysis of teacher feedback has predominantly been done in experimental settings or using a limited sample of grade transcripts with a feedback analysis that remained mostly descriptive (Demoulin and Daniel, 2013). Our first contribution is to build a statistical model to shed light on teachers gender-differentiated feedback in a data-driven manner. We perform a textual analysis of comprehensive and non-experimental data and propose a synthetic measure of teachers' gender-differentiated feedback based on the written feedback provided to students. Using text mining techniques on the universe of Grade 12

[1] See Alan et al. (2018); Carlana (2019); Copur-Gencturk et al. (2020) for recent references documenting teachers' stereotyped beliefs.

students' math transcripts over the period 2012–2017, we build a statistical model that predicts students' gender relying on the words used by math teachers. For each teacher, the share of correctly predicted observations, also known as the accuracy of the model, is our measure of teacher gender-differentiated vocabulary (GDV hereafter). The more a teacher makes use of female predictors to qualify its female students' work and male predictors to qualify its male students' work, the better the predictive accuracy, and the stronger the teacher gender differentiation. To get a sense of how different the vocabulary used by high GDV teachers is relative to gender-neutral ones, we further classify gender predictors into positive *versus* negative, and managerial *versus* competence-related categories.[2]

The extensive analysis of the vocabulary used in transcripts and the synthetic measure of gender-differentiated feedback proposed in this paper allow us to describe the feedback patterns used by French Grade 12 math teachers. We find that, on average, math teachers differentiate their vocabulary according to their students' gender, and document a large variation in the distribution of the teacher GDV measure. Another finding is that math teachers are significantly more differentiating their vocabulary than teachers in humanities. Using the feedback provided by teachers in those subjects and applying the same estimation procedure yields lower predictive accuracies, meaning that the words used by these teachers are more gender-neutral than those of math teachers.

The classification of the best gender predictors along the four above-mentioned dimensions reveals marked gender differences in the vocabulary used by math teachers to describe the work of equally able students. The first striking fact concerns the relative proportion of positive and negative feedback received by each gender. Out of the 30 best female predictors, two-thirds are classified as positive and mostly relating to their behavioral skills and efforts. The reverse is observed for male predictors, that relate to negative managerial aspects in two-thirds of cases. Positive male predictors however praise their intellectual skills. Overall, math teachers insist more on positive managerial aspects and encourage the efforts provided by their female students, while equally performing males are both more criticized for their unruly behavior and praised for their intellectual skills.

The second contribution of this paper is to relate our teachers' GDV measure to students' academic performance and enrollment outcomes. We investigate the extent to which exposure to a high GDV teacher affects students' percentile rank at the national *baccalauréat* exam in different subjects. We also investigate the impact on their higher education choices by looking at

[2] The words classified as "managerial" refer to the student's attitude in class as well as the efforts and actions undertaken for the subject, while those classified as "competence" relate to math concepts and the school environment or to the students' intellectual ability (Morgan, 2001).

the rank-order lists that students submit when applying to higher education programs. Finally, we track students the year following their high school graduation and look at whether high GDV teachers also influence enrollment decisions. Our identification strategy exploits the within high school variation in teacher GDV, and relies on the fact that teachers' allocation to the different classes is almost as good as random conditional on some observable characteristics.

We find that being assigned to a teacher with a one standard deviation higher GDV increases students performance at the math *baccalauréat* national exam by 0.34 percentile rank on average (from a baseline of 50), the effect being slightly larger for female students (+0.37 percentile rank against +0.31 for males). These coefficients capture the effect of moving from an average GDV-teacher to a teacher from the 86^{th} percentile of the GDV distribution. Even though moderate on average, the effect on math percentile rank at *baccalauréat* increases with teachers' GDV, and even more so for female students. Compared to those who are exposed to teachers from the bottom 10 percent of the teacher-GDV distribution, female students exposed to teachers from the 4^{th} decile or above see their percentile rank in math increase by up to 2 percentile ranks. These results suggest that teachers who are more likely to differentiate their vocabulary by praising females mostly for their efforts and their in-class behavior enhance their performance.

Most of the effects on the type of programs students rank when applying to higher education or on the type of programs where they eventually matriculate in the following year are small in magnitude and not statistically significant. Being exposed to a one standard deviation higher teacher GDV leaves female students' likelihood of top-ranking a STEM program unchanged, and slightly decreases that of male students, by 0.27 percentage point, significant at the 5 percent level. The effects on matriculation in the year following high school graduation are close to those on top-ranked programs.

Related literature. Our paper speaks to different strands of the literature. First, it relates to the broader literature on the origins of women's underrepresentation in male-dominated fields. While this literature initially focused on gender differences in math performance, whereby males are said to outperform females in math test scores, a consensus seems to emerge on the fact that those test score gaps are quite limited (Lindberg et al., 2010), and are themselves influenced by social and cultural norms, as more egalitarian countries usually record lower gender performance gaps (Guiso et al., 2008; Breda et al., 2018). Cultural and social norms are now privileged as the most plausible explanation for such gender gaps. Social norms, as conveyed by parents, peers and teachers are major determinants of young females' math perceptions, school performance and future educational choices.

Parents' and teachers' positive attitudes towards math tend to spill over to their children and pupils, who embrace a growth mindset, i.e. who believe that one can succeed by exerting effort (see Gunderson et al. (2012) for a review). Such a positive attitude is associated with a higher math performance and increased graduation in STEM fields (Cheng and Kopotic, 2017). Math teachers and parents working in STEM occupations may also act as role models, by breaking stereotypes and raising interest for science-related careers (Lim and Meer, 2017; Eble and Hu, 2019a; Cheng and Kopotic, 2017), even though they sometimes engage in different teaching practices or behavior with female and male students. For instance, Oguzoglu and Ozbeklik (2016) find that fathers in STEM occupations are much less likely to transmit STEM-specific tastes to their daughters when they also have a son, while Lavy and Sand (2018) and Terrier (2020) find that math teachers' grading behavior is biased in favor of female students. Peers also play a role in the perpetuation of gendered-beliefs and stereotypes. Classmates are a vector of their parents' beliefs. Exposure to peers whose parents believe that males are better than females at learning math increases a child's likelihood of believing it as well, harms females' but improves males' performance in math (Eble and Hu, 2019b). Peers may also convey such social norms by penalizing classmates who deviate from the norm (SkoǨajić et al., 2020).

Our paper also relates to the growing research in social sciences using text as data (Gentzkow et al., 2019) to uncover patterns of biases and discrimination. In an experimental setting where questions are randomly asked by a fictitious male or female account on a math forum, Borhen et al. (2018) show that the answers to questions posted by female accounts contain significantly more positive *and* negative words and opinions than answers to males. Wu (2018) analyses threads from the Econ Job Rumor Forum and shows that the language used to qualify women in academia is substantially different from that of males. When women are mentioned in a thread, the discourse is significantly less related to professional matters and focuses more on personal information or physical appearance. Koffi (2020) uses bibliometric data to investigate gender biases in citation patterns of Economics scholars. She shows that omitted papers, i.e. the ones that are relevant and should be cited, are 15 percent to 30 percent more likely to be female-authored than male-authored, and that this omission bias is twice as large in more theoretical fields involving mathematical economics. Ningrum et al. (2020) perform a textual analysis of job advertisement and find explicit discrimination against females in the hiring process.

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

2 Institutional Background

2.1 The French Secondary Education System

In France, the secondary education system consists of seven years of schooling, divided into four years common to all students and taught in middle schools (*collège*, Grade 6 to 9), and three years of high school (*lycée*, Grade 10 to 12), either delivering a vocational or a general and technological training. Both middle and high school curricula end with a national examination. At the end of middle school, students take the *Diplôme National du Brevet* (DNB), which tests their knowledge and skills in math, French and history and geography. At the end of Grade 11, high school students take the anticipated *baccalauréat* examinations, which include oral and written tests in french, as well as in history and geography for science major students. Students are tested in the remaining subjects at the end of Grade 12. Only students holding the *baccalauréat* can enter the higher education system.

In general and technological high schools, after a common *Seconde générale et technologique* year (Grade 10), students are tracked into a general (80 percent of students) or a technological curriculum (20 percent of students). General track students further specialize by choosing their major, when entering Grade 11, and their elective course, when entering Grade 12. Students tend to specialize according to both their comparative advantage and their preferences, which leads to marked gender patterns in major and elective course choices. While females are slightly underrepresented among science major students (they represented 47 percent of the science major students in 2018), the economics and humanities majors are largely female-dominated: in 2018, 60 percent of economics major students and 80 percent of humanities major students were females (MENJ-MESRI 2019). These gender patterns in major choice are further reinforced by the choice of elective courses. The differences are particularly striking when focusing on science major students. Female students are largely overrepresented in the earth and life science elective, where they represent 63 percent of students, against only 30 percent in computer sciences and 15 percent in engineering. The proportions of female students in math and physics-chemistry electives are more balanced (42.5 percent and 48 percent respectively).

The gender segregation within French high schools is however limited beyond the segregation induced by the choice of an elective course. The composition of each class is determined by high school principals who, while taking into account the students' electives when defining the classes, also declare putting gender diversity on top of their priority list (Cnesco, 2015). Most principals also declare valuing some heterogeneity in terms of students' academic achievement level but, unlike for gender, the academic stratification within high schools remains substantial. Ly and Riegert (2015) have looked at the determinants of the within high school segregation and found that grouping students according to their elective courses accounts for two-thirds of the observed social and academic segregation.

2.2 College Application and Enrollment

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

The undergraduate programs students can apply to fall into two broad categories, with, on the one hand, university programs, which are mostly non-selective and open to all high school graduates, and, on the other hand, selective programs.[3] The latter include three different types of curricula, which have a strict academic stratification: two-year undergraduate vocational and technical programs (*sections de techniciens supérieurs* and *instituts universitaires de technologie*), undergraduate management and engineering schools, and the two-year elite *classes préparatoires aux grandes écoles* (CPGE). The CPGE prepare students to the entry exam to the most prestigious French colleges (the *grandes écoles*) in science, business, or humanities. Science CPGE are further specialized into three main categories: MP (math and physics), PC (physics

[3] Since the 2018 reform of the application procedure that changed Admission Post-Bac into Parcoursup, universities are now allowed to select students according to their past academic performance but our study does not cover the post 2018 period.

and chemistry) and BCPST (biology), while business CPGE can be distinguished between sciences, economics and technological programs, and humanities CPGE are composed of classics and social sciences programs.

Until 2017, the college admission procedure was centralized through the *Admissions Post-Bac* (APB) online platform for most undergraduate programs. The main round of the student assignment mechanism relied on a procedure that was close to the college-proposing deferred acceptance algorithm (Gale and Shapley, 1962; Roth, 1982). Students were invited to submit a rank-order list of programs that could include up to 36 choices, with a maximum of 12 choices per type of program (University program, STS, CPGE, etc.). After the list's submission deadline, students were ranked by the different programs. For selective programs, the ranking was based on their Grade 11 and 12's academic records. The grades obtained in different subjects as well as teachers' written feedback played a crucial role in students' ranking. For non-selective programs, students were ordered according to some priority rules, based on their catchment area and the program's rank in the student's list.

3 Data and Summary Statistics

This section details the different data sources that we use to build our measure of teacher GDV and to quantify its impact on students' outcomes (section 3.1). We also present summary statistics on the sample of Grade 12 students and on their math teachers (Section 3.2).

3.1 Data Sources

We use three main administrative databases: the college application data for six cohorts of Grade 12 students (2012-2017) as well as information on their math teachers collected via the APB platform; the higher education enrollment data; and the data of the two main national exams (*Diplôme national du brevet*, DNB, and *baccalauréat*).

APB data. Our primary source of information is the comprehensive application data from the APB platform over the period 2012–2017. A substantial amount of information is collected by this platform during the application process. We first use the students' digitalized academic records to retrieve teachers' feedback on all the subjects taken by students during Grade 11 (three trimesters) and Grade 12 (two trimesters only). This is the main input used to build our measure of teacher GDV. Teachers and students are uniquely identified in the data, which enables us to link these transcripts to students' and teachers' characteristics contained in a separate APB file. Along with basic sociodemographic information on the students (gender, place and date of birth, parental socio-economic status, etc...), the APB data provide detailed information on the schooling trajectories of the students in high school (school track, major and elective choices, etc...). The data also allow us to determine the teachers' gender, the subject they teach and whether they are the head teacher of the class.

Another interesting feature of these data is that they keep a record of the final rank-order list of programs submitted by each participating student. For each year, we also know the matching outcome, i.e. the program to which each student was admitted along with the students' acceptance decision (acceptance, conditional acceptance or rejection).

School performance data. We use the OCEAN database, managed by the French Ministry of Education, for the grades obtained in two national examinations: the *diplôme national du brevet* (DNB), taken at the end of Grade 9, and the *baccalauréat*, taken at the end of Grade 12. We use the former to control for the students' past academic performance in the estimation procedure, while the latter is used as an outcome for student's performance at the end of high school. For that purpose, and to make grades comparable across years, we transform the initial grades ranging between 0 and 20 into percentile ranks, where 0 and 100 are respectively the ranks for the lowest and the highest performing students. It is worth mentioning that both exams are "blind" tests, i.e. they are anonymously and externally graded.

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

Sample restrictions. Given that the focus of this study is on the impact of math teachers' vocabulary, we restrict our sample to Grade 12 students enrolled in the science track as they are the ones interacting most frequently with their math teachers, relative to the humanities or the social sciences tracks.[4] These students are also the most likely to choose a science college major after graduation and may therefore be more responsive to the math teacher's feedback. We then drop students for whom the math teacher's identifier or the grade transcript is missing, which represents about 50 percent of Grade 12 students from the science track in 2012 and goes down to 15 percent in 2017 (Table 1). In the vast majority of cases (between 70 and 95 percent of missing observations), teachers' identifiers and grade transcripts are missing because the entire high school is not reporting its students' grade automatically on the APB platform. Dropping those observations therefore amounts to dropping entire high schools and is not a threat for the internal validity of our analysis.[5] We finally restrict our sample to high schools having at least two science track classes – as our identification strategy relies on a within-school comparison of students (see section 4), and to teachers having taught at least two classes over the period. These restrictions lead us to exclude between 6 and 20 percent of students. Once those restrictions have been applied and depending on the year considered, the sample of analysis consists of 40 to 75 percent of Grade 12 students from the science track, for a total of approximately 700,000 observations.

3.2 Summary Statistics

Students' characteristics. Table 2 provides summary statistics of Grade 12 science track students' characteristics for the whole sample of analysis and separately for male and female students. Students are eighteen years old on average and mostly come from a high (43 percent) or a medium high socio-economic background (16 percent).[6] Female students are slightly underrepresented in the science track as they account for 47 percent of science track students but 54 percent of all general Grade 12 students (MENJS-MESRI, 2018). Turning to elective courses, we note striking gender differences. Half of female students opt for the earth and life science elective against only one fourth of males. Female students are also underrepresented in the math (19 percent against 27 percent) and engineering and computer sciences electives

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

[5] It might, however, affect the external validity of our analysis. Table C5 in Appendix C shows the OLS coefficients of a dummy indicating whether the high school has all grade transcripts missing regressed on the high school's average characteristics. High schools with a higher share of female and free lunch students are more likely to be reporting the grade transcripts. Reassuringly, the relative performance of female vs. male students at the math DNB examinations only marginally affects the probability of not reporting grade transcripts.

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

(6 percent against 20 percent). Another noticeable difference between male and female students relates to their past academic performance, as measured by the national percentile rank at the DNB math exam. Males' average rank is approximately 4 points above that of females. Figure 1 further shows that males are largely overrepresented in the top quartile of math performance: they make 58 percent of the top math achievers. On the other hand, females outperform males in French at both the DNB and *baccalauréat* exams, with an average percentile rank that is 10 points higher than that of their male peers. Both imbalances, in terms of elective choices and past school performance, are accounted for in our identification strategy and during the estimation procedure.

Students' higher education enrollment. The gender differences observed for elective courses are also found when looking at enrollment in the higher education system, as shown in Figure 2. Both males and females enroll in a scientific program or in medicine in the same proportion: 54 percent of females and 58 percent of males opt for a STEM program (be it selective or not) or medicine. About one-fifth does not enroll in higher education and the remaining choose a non-scientific program (selective or not). However, we do observe substantial gender segregation within scientific subjects. The proportion of females enrolling in any STEM program is much lower than that of males. Only 11 percent of female students choose a selective STEM program against 21 percent for their male counterparts. Those proportions are respectively 16 percent and 26 percent for enrollment in a math, physics, chemistry or computer science program at university. Females rather enroll in a medicine program: 27 percent of Grade 12 females from the science track opt for such program against 11 percent of males only.

Math teachers' characteristics. Table 3 reports some descriptive statistics for the sample of math teachers in Grade 12 Science track. There are 6,772 math teachers in the sample, 58 percent of which are males. A little more than half of them have been the head teacher of a class at least once over the period covered by our data. Those teachers are likely to have a stronger influence on students' performance and enrollment behavior as they counsel students on top of teaching them. Each teacher is in charge of only one Grade 12 class from the science track on average each year, with an average class-size of 28 students (90 percent of teachers teach only one Grade 12 science track class per year). Teachers appear almost four times in the sample, meaning that we have on average four teacher×classroom observations, which is crucial for the reliability of our GDV measure (see section 4). Finally, the average feedback is made of 7.4 words, with large variability in feedback length across teachers.

4 Empirical Strategy

The first part of this section explains the estimation procedure used to measure teachers' gender differentiated vocabulary (GDV) (section 4.1). The second part presents the identification strategy to estimate the impact of teacher GDV on students' outcomes (section 4.2).

4.1 Measuring Teachers' Gender Differentiated Vocabulary (GDV)

The measure of teacher GDV proposed in this paper leverages the rich data on teachers' feedback provided to students three times a year in their Grade 12 academic records. This personalized feedback reflects the teachers' perception of the students' performance, work, and behavior in class throughout the year. To investigate whether the words used to characterize the students' work, behavior and ability differ by gender, we build a model that predicts students' gender based on the words used in the teachers' feedback. Using machine learning techniques, we estimate a model on the sample of Grade 12 science major students. We then use this fitted model to compute a measure of gender differentiated vocabulary for each teacher based on her observations only, controlling for class-level gender imbalances in students' prior academic performance. The different estimation steps are presented below while the detailed procedure can be found in Appendix A.

Data preparation. Converting the corpus of teacher feedback into a statistical database is done in two steps. First, we rely on text mining techniques to replace each word by its root and hence ensure its gender neutrality (Gentzkow et al., 2019). Second, the corpus of teachers' feedback is turned into a matrix that contains one row per feedback and W_n columns, where W_n is the number of distinct words appearing in the corpus. Each of these columns is a dummy that takes the value one if the considered word appears in the student's feedback, and zero otherwise.[7] Finally, we classify words into one of the four following categories, inspired from the sociology and psychology literature: positive (resp. negative) competence-related aspects and positive (resp. negative) managerial aspects. 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 labeled neutral or unclassified.[8] This classification is not directly used to build the measure of

[7] Increasing the flexibility of the model by adding interactions between words (*bigrams*) does not improve the predictive accuracy of the model. We therefore stick to the simplest model specification, involving single words (*unigrams*) only.

[8] The classification of words is detailed in Tables A2 and A3 of Appendix A

GDV but is necessary to characterize along which dimensions the vocabulary used by teachers differ.[9]

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

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

Our objective is to find the set of α coefficients that minimize a penalized version of the log likelihood $\ln(L(\alpha))$ associated to Model 1, where λ is the regularization parameter:[10]

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

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:

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

On average, our model correctly predicts students' gender in 60 percent of cases, and performs better at predicting males' gender than females', as 67 percent of males are correctly classified against 53 percent for females.

Teachers' gender differentiated vocabulary (GDV). We define each teacher j's GDV for class c as the predictive accuracy of Model 1 fitted on his or her students only, where the predictive accuracy is the share of the teacher's students whose gender is correctly predicted by the model.[11]

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 terms of math performance. To

[9] We attempted to build these categories in a data-driven manner using bi-term topic models tailored for short texts, but these models performed poorly on our data. Our data is indeed quite specific in that texts are very short, with an average number of tokens equal to 7, the overall vocabulary is quite limited ($\simeq$ 1,600 words) with little variation in the topics used as they all relate to academic performance and behavior. We therefore faced the typical challenges inherent to such short texts: the generated topics gathered inconsistent words (*trivial topics*) and the different topics were highly similar with a lot of words in common (*repetitive topics*, see Wu et al. (2020) for a discussion on those issues.)

[10] We select the regularization parameter λ using a 10-fold cross-validation procedure, and pick the highest λ value such that the error of prediction lies within one standard deviation of the minimal error (Hastie et al., 2009).

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

alleviate this concern, we predict teacher j's GDV for class c on a balanced subsample of students. More specifically, for each teacher, we undersample as many males and females from each quartile of prior math ability, so that we have 50 percent of male and female students at each ability level.[12] The quality of the prediction could also be influenced by the feedback's length which varies by teacher (Table 3). For lengthy feedback, defined as the ones with an above-median length, we randomly sample six words, i.e. the median number of words.

Another related concern is that teacher j's GDV estimated for class c could capture some unobserved class-specific gender differences in behavior or performance. To rule out this possibility, we compute an alternative measure that we call the *leave-one-out* GDV, defined as the average of teacher j's GDV over all the classes she taught during the study period, excluding class c. Our two measures are defined as follows:

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

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

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

where N_j is the number of classes that teacher j taught throughout the period under study.[13]

Both GDV measures theoretically lie between 0 (the model systematically misclassifies females as males and males as females) and 100 (all students are assigned their actual gender). The higher the accuracy for a given teacher, the better we can, on average, recover his or her students' gender based on the words she uses in her feedback, and hence the stronger the gender differentiation in the vocabulary that she uses in her assessments. A model that randomly assigns each student a gender with probability 0.5 would achieve an accuracy of 50 percent, meaning that our model predicts gender better than random guessing for all teachers whose accuracy is above 50 percent. An accuracy below 50 percent is possible in our setting given that the prediction is done on small samples at the teacher level. However accuracies are averaged over 100 estimations to limit such random fluctuations, and the *leave-one-out* GDV is itself an average of several accuracies, therefore reducing the noise inherent to the measure.

[12] Prior ability in math is proxied by the percentile rank in math at the DNB exam.

[13] In practice, both GDV measures are computed as the average of GDV and *leave-one-out* GDV estimated on 100 random balanced subsamples of teacher j's students.

4.2 Identification Strategy

The second objective of this paper is to establish a causal link between the teacher GDV measure presented above and students' performance and enrollment outcomes. Our identification strategy relies on the comparison of students enrolled in the same high school but exposed to math teachers with different levels of GDV. More specifically, we exploit the within high school variation in teacher GDV and estimate the following equation, controlling for year and elective courses fixed effects:

$$Y_{isjet} = \alpha + \beta_1 GDV_{j\backslash c} + \gamma_s + \eta_e + \delta_t + \epsilon_{isjet} \tag{6}$$

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. $GDV_{j\backslash c}$ is teacher j's standardized GDV measure and is class-specific as we use the *leave-one-out* GDV described in Equation 5. γ_s is a set of high school dummies and η_e and δ_t are elective and year fixed effects respectively. The coefficient of interest is β_1, which measures how a student's outcome is affected by being assigned a teacher with a one standard deviation higher GDV. The standard errors are robust and clustered at the teacher level.[14]

The validity of our identification strategy requires that teacher GDV is not systematically correlated with students' characteristics. We formally test this in Section 6.

5 Math Gender-Differentiated Feeback

In this section, we first provide an overall picture of the gender differentiation of the vocabulary used by math teachers (section 5.1), before turning to the analysis of our GDV measure (Section 5.2).

5.1 Descriptive Statistics on Math Feedback by Gender

This section aims at providing a comprehensive picture of what a math feedback looks like for a Grade 12 male or female student. We first provide statistics on the distribution of word counts of math feedback overall and by type of feedback, and then provide statistics aggregated at the teacher level.

[14] Although standard errors should be bootstrapped to account for prediction error, we do not implement this correction due to computational limitations.

Students' math feedback. Panel (a) of Figure 3 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.[15] The first finding is that female students tend to receive slightly shorter feedback: while the median feedback addressed to male students is made of seven content words, the median female student feedback contains only six content words. The upper tails of the feedback length distributions are similar across gender, with 25 percent of feedback containing more than nine content words, but the lower tails differ slightly, with 25 percent of feedback containing less than five words for male students, and less than four words for female students.

Those summary statistics are then broken down according to the four dimensions mentioned in Section 4: managerial vs. competence-related feedback and positive vs. negative feedback. The summary statistics along the positive vs. negative dimensions do not reveal significant differences between male and female students. For both genders, the median feedback contains two positive words and zero negative word, and 25 percent of the students receive more than three positive words and 1 negative word. The managerial and competence-related dimensions highlight stronger gender patterns at the top of the distribution only. The median feedback addressed to male and female students contains one competence-related word and two managerial-related words. However, while 10 percent of male students receive more than four competence words, this is the case of only eight percent of female students.

Math teachers' feedback. In Figure 4, we take a teacher perpective to dig into the gender differentiation of the type of feedback provided. For each teacher, we compute the share of managerial or competence-related feedback separately by gender, and plot the distributions of teachers' gender gaps in the shares of managerial and competence-related feedback in Panel (a).[16] While 81 percent of the teachers put a stronger emphasis on managerial matters in the feedback addressed to females compared to males, only 18 percent of the teachers insist more on competence-related aspects when they assess a female student's work and performance. We further characterize these managerial and competence-related feedback along the positive and negative dimensions. Panel (b) shows the differential propensity to address a positive feedback across gender: 80 percent of the teachers tend to be more positive for females than for males on managerial-related issues, while this is the case of only 20 percent of the teachers for competence-related matters. The propensity to be negative on managerial-related feedback is

[15] For simplicity, we define as content words all words that are not stop words (cf Section A).

[16] The shares are computed on the subset of words that could be classified either as competence-related or managerial.

more balanced (44 percent of the teachers are more negative with females), while the weight given to negative feedback on competence-related matters is more pronounced for females, with 77 percent of the teachers that are relatively more negative with them.

5.2 Math Teachers' Gender Differentiated Vocabulary (GDV)

The naive analysis performed in the previous section reveals clear gender patterns in math teachers' feedback, but the observed differences are potentially resulting from genuine gender differences in student ability and attitudes towards math. We now turn to the analysis of our GDV measure estimated with the methodology described in Section 4, which is explicitly designed to isolate feedback patterns that are independent from students' own characteristics.

Distribution of teachers' GDV. Figure 5 shows the density as well as the cumulative distribution of the GDV and *leave-one-out* GDV measures separately. It provides evidence of the existence of a correlation between students' gender and the feedback received in math, controlling for the students' prior ability level in math. Our model predicts gender better than random for 90 percent of math teachers if we consider the GDV measure and for over 95 percent with the *leave-one-out* GDV.[17] It correctly predicts the gender of 61 percent of the students for the median teacher.[18] When breaking down the GDV distributions by teachers' gender, we note that, on average, female math teachers differentiate their vocabulary slightly more than their male colleagues (see Figure 6). Another finding is that using a gender-specific vocabulary is something quite specific to math-intensive subjects. We replicate the teacher GDV estimation procedure on the feedback provided to our sample of Grade 12 students in the following core subjects: physics & chemistry, biology, philosophy and modern language 1 and 2. Figure 7 shows that the *leave-one-out* GDV distribution for humanities-related subjects is shifted to the left compared to science-related subjects. This suggests that teachers in philosophy and modern languages are, on average, less inclined to use a gender-specific vocabulary in their feedback compared to math, physics and chemistry teachers, while biology teachers are somewhere in-between. Philosophy is a particularly relevant point of comparison as its teacher-gender composition is close to that of math-intensive subjects (62% of males, see Table B). Yet, philosophy teachers seem to use a more-gender neutral vocabulary.

[17] In total, only 4 percent of teachers have a leave-one-out GDV below 50 percent, the vast majority of which falls between 44 percent and 50 percent. This is mostly explained by the fact that those teachers are observed 3 times on average, compared to 4 times for other teachers. Their leave-one-out GDV is therefore slightly noisier.

[18] For comparison, when predicting whether the student's performance at the math DNB examinations corresponds to the top or bottom half performance in his class, the model achieves a median predictive accuracy of 65.8 percent.

Qualitative analysis of the best predictors. A high degree of gender differentiation in the vocabulary used, as measured by a high GDV measure, can reflect different teachers' attitudes. A gender differentiated feedback may be the expression of gender stereotypes regarding students' ability in math, but may also reveal an effort from the teacher to adapt her feedback to the different student profiles she perceives. In order to better understand to what extent the gender differentiation of the vocabulary used by math teachers reflects either of the two attitudes, we analyze the best gender predictors and classify them using the four categories defined in Section 4 to ease the comparison between male and female students' feedback.

The analysis of the best predictors of each gender points towards marked differences in the qualifiers used by teachers. Figure 8 reports the odds ratios derived from the estimation of the model described by Equation 1, for the top 10 predictors of each gender.A feedback mentioning the student's confidence level, his propensity of getting discouraged or his cheerful aspect ("smiley") is between 1.8 and 2.3 times more likely to be directed to a female than to a male student, relative to a feedback that does not mentioning it. Teachers are also more likely to mention that female students are stressed or panicked, and to insist on their exemplary conduct ("exemplary", "studious"). On the other hand, a feedback describing the student as childish ("childish", "has fun"), insisting on the need for careful handwriting, or praising the student's curiosity and intuitions is about 2.5 times more likely to be that of a male rather than that of a female student.

Figure 9 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), and where the colors refer to competence, managerial or unclassified words. The first striking feature of this graph is the relative proportions of positive versus negative type of feedback by gender. Among the top 30 male predictors, only 5 correspond to a positive feedback, while roughly two thirds of the best female predictors can be considered as positive. Most interestingly, conditional on being positive, the best male predictors almost all qualify the student's competence-related aspects ("curious", " idea", "interest", "intuition"), while nearly all of the best female predictors qualify managerial aspects ("irreprochable", "willingness", "persevere"). On the other hand, more than 80 percent of the best male predictors can be classified as negative and the vast majority refer to a disruptive behavior ("has fun", "childish") or to a neglected work-effort ("waste", "superficial").

Finally, these results are confirmed when we consider all gender predictors. Figure 10 first shows the proportions of negative, positive and neutral feedback conditional on having a competence-related feedback (left-hand side of the graph) or a managerial one (right-hand-side).

Among the predictors for being a female student that can be classified as competence-related, only 20 percent correspond to a positive feedback against 39 percent for male predictors, while 17 percent versus 12 percent are negative, the rest being neutral. Symmetrically, among the female predictors that can be classified as managerial, 44 percent correspond to a positive feedback against 27 percent for males. The latter get a much larger share of negative feedback: as much as 45 percent of managerial male predictors are negative while this proportion is only 29 percent for females.

Turning to the proportions of competence, managerial and neutral words conditional on having a positive or a negative feedback, we see that conditional on being positive, top female predictors qualify their competence-related skills in only 16 percent of cases against 40 percent for their male counterparts. Regarding the breakdown of negative predictors, 40 percent of negative male predictors relate to managerial matters against 33 percent for female predictors, and those proportions are respectively 16 percent and 9 percent for negative competence-related predictors.

Vocabulary used by teachers' decile of GDV. The last thing we investigate is whether teachers with varying degrees of GDV differ from each other, by comparing the teachers' gender gaps in the share of positive words among competence and managerial-related feedback by decile of GDV (see Figure 11). Panel (a) plots the absolute values of the teachers' gender gaps and Panel (b) displays the share of teachers having a gender gap in favour of females, separately for competence and managerial-related feedback. The gender gaps in the share of positive words increase at a growing pace with GDV deciles, indicating that teachers with a higher GDV tend to provide relatively more positive feedback to female students, and even more so as the GDV decile is high. This is true for both managerial-related and competence-related feedback, with a gender gap that goes from 6 to 7 percentage points in the lower GDV deciles up to 10 points in the 10th decile. In line with the findings resulting from the analysis of the gender predictors, Panel (b) reveals that higher GDV teachers are overwhelmingly relatively more positive with females on managerial-related feedback, and more negative with females on competence-related feedback.

Taken together, these descriptive statistics indicate that teachers do use a differentiated vocabulary for their male and female students. They seem to insist more on positive managerial aspects and to encourage the efforts provided by their female students, while equally performing males are both more criticized for their unruly behavior and praised for their intellectual skills. In the following section, we investigate to what extent this gender differentiation in the feedback provided affects students' performance and future enrollment outcomes.

6 Impact of Teachers Gender-Differentiated Vocabulary on Students' Outcomes

After having documented differences in Grade 12 math teachers' gendered vocabulary, we turn to the impact of teachers' GDV on students' outcomes. We first perform a series of statistical tests aimed at validating our empirical strategy (Section 6.1). We then discuss the impact of teachers' GDV on students' academic performance and on their higher education choices and enrollment the year following high school graduation (Section 6.2). We show that our results are robust to a series of alternative specifications.

6.1 Validy of the Empirical Strategy

6.1.1 Exogeneity Assumption

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

Balancing tests between students' characteristics and teacher GDV. Table 4 reports the coefficients from a regression of the teachers' standardized *leave-one-out* GDV, defined at the class level, on students' socio-economic characteristics and baseline academic performance, along with a set of high school, year and elective courses fixed effects. The table shows that teacher GDV is not systematically correlated with students' observable characteristics. Out of the twelve characteristics included in the regression, only the age variable is significant at the 10 percent level, and the magnitude of the coefficient is very low.[19] When jointly tested, we cannot reject the null hypothesis that baseline variables are not correlated with the *leave-one-out* GDV measure, thus providing evidence of teachers' random allocation conditional on high school, elective and year fixed-effects.

Random allocation of students. To check that students are randomly allocated to teachers within a given high school, elective course and for a given year, we follow Lavy and Sand (2018) by performing a series of Pearson's Chi-square tests of independence. For each of the 27,668 high

[19] The coefficient can be interpreted as follows: Increasing teacher GDV by one standard deviation is associated with a 0.0043 year decrease in student's age.

schools, elective course, and year unique combinations, we tabulate math teachers' identifiers with each of the students' baseline characteristics and test for independence.[20] Table 5 reports the percentage of p-values below the nominal values of 0.05 and 0.01. Except for the female dummy, we find that the empirical p-values are close to the nominal values (between 4.5 percent and 8 percent of p-values are below nominal levels, i.e. significant). For the female dummy, the empirical p-value is 11 percent, suggesting that in 11 percent of high school×elective×year combinations, we cannot exclude the non-random assignment of female students to classes at the 95 percent-level.

Taken together, the tests performed in this section suggest that in a given high school, elective course and for a given year, students are close to being randomly allocated to classes. To ensure that the results presented in Section 6.2 are not driven by the slight gender imbalances, equation 6 is also estimated with the average proportion of females in the class as an additional control, as well as the full set of students' baseline characteristics.

6.1.2 Reverse Causality

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

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

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

Third, looking at the distributions of leave-one-out GDV measures estimated for other subjects further highlights the specific nature of science-related subjects, for which students' gender is better predicted on average than for humanities-related subjects (see Figure 7). Students' gender is correctly predicted in 59 percent of cases for humanities-related subjects, against 61 percent for math or physics and chemistry on average. Science-teachers leave-one-out GDV distributions are shifted to the right compared to that of humanities-related subjects. This

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

suggests that on average, for a given class, science teachers differentiate more their vocabulary by gender. Our measure therefore captures differences that go beyond class-specific characteristics.

Finally, the fact that teachers' GDV is computed for multiple years and classes offers the opportunity to measure the persistence of teachers' GDV across classes and years, to ensure that we are not only capturing noise. The correlation between the teacher's GDV and leave-one-out GDV, i.e. the correlation between a given GDV and its average computed in other years×classes, is 0.184 and is statistically significantly different from zero.[21] It is worth noting that as we are correlating several GDV measured with error because of the small sample size used to make the prediction at the class level, this correlation suffers from an attenuation bias. As a comparison, in the teacher value-added literature, the within-teacher correlation is usually around 0.3 (Chetty et al., 2014).

Overall, we are confident that our GDV measure captures some persistence in the teachers' behaviour.

6.2 Impact of Teacher GDV on Performance and Enrollment

Results on Academic Performance. Panel A of Table 6 reports the estimated impact of teacher's *leave-one-out* GDV on students' percentile rank at the math *baccaularéat* exam, obtained from the estimation of Equation 6. As a placebo test, the table also displays the teacher GDV's impact on the students' percentile rank at the philosophy *baccaularéat* exam.

A one-standard deviation increase in the math teacher GDV raises math performance at *baccalauréat* by 0.34 percentile rank on average, significant at the one percent level. In other words, this is the effect of moving from an average teacher in terms of GDV to a teacher from the 86^{th} percentile of the GDV distribution. The effect is slightly larger for females, whose percentile rank increases by 0.37 when exposed to a one standard deviation higher GDV teacher, against 0.31 for males, but the difference is not statistically significant. As expected, the math teacher's GDV has no effect on the percentile rank in philosophy, which is evidence that the GDV measure affects students' outcomes through the considered teacher only, and does not capture class-specific effects. Those results hold when we control for students' baseline characteristics, for the share of females in the class or for the average GDV measured in other subjects for students from the same class (see Table D6 in the Appendix).

Even though moderate, these average effects hide heterogeneous responses that depend on the degree of gender differentiation of the teacher's vocabulary. Instead of including the

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

teacher GDV linearly in the equation, we explore the intensity of the treatment by regressing the students' outcomes on a set of GDV deciles. The first (last) decile corresponds to the bottom (top) 10 percent of the math teachers' leave-one-out GDV distribution. Figure 12 plots the deciles' coefficients along with their 95 percent confidence intervals, separately for males and females. The figure shows that the higher the teacher GDV, the higher the effect on math performance, and even more so for female students. Compared to females exposed to the bottom 10 percent of teachers in terms of GDV, those exposed to teachers from the 4^{th} decile or above see their percentile rank at *baccalauréat* in math increase by 1.5 to 2 ranks on average, significant at the 5 percent level. The effect for females whose math teacher's GDV falls below the 4^{th} decile ranges between 0.7 and 1.5 and is also statistically significant at the 5 percent level. This trend is also observed for males, but to a lower extent. Deciles' coefficients range between 0 and 1.5 percentile ranks, and are always below that of their female counterparts.

We also explored various dimensions of heterogeneity such as students' prior math performance or socio-economic status but did not find evidence or differentiated effects. Results are reported in Appendix D.

Results on choice lists and matriculation in the following year. We then look at the effects associated to a one standard deviation increase in teacher GDV on the type of top-ranked STEM programs in students' final rank-order lists (Figure 13 and Table 6, Panel B), and on their actual matriculation in the year following high school graduation (Figure 14 and Table 6, Panel C). Being exposed to a teacher with a one standard deviation higher GDV does not alter females and males' STEM first choice for higher education programs nor does it affect their enrollment outcomes in the year following high school graduation. If anything, males are only marginally less likely to enroll in a selective program in the following year (-0.22 percentage point) or in a vocational program (-0.10 percentage point), which represents a 1 percent and a 2.25 percent decrease with respect to the baseline proportions of males enrolling in such programs, as reported in Figure 2. We do not document any heterogeneous effects by deciles of teacher GDV, by student's prior math performance or by student's socio-economic status.

7 Mechanisms

In this section, we explore the mechanisms that could potentially drive the effects of math teachers' GDV on math performance at *baccalauréat*. We first look at whether teachers using a gender differentiated vocabulary are also encouraging females by overgrading them relative to males (section 7.1). We then compute a measure of teacher quality to investigate whether math teachers with a higher GDV are also better teachers (section 7.2). Finally, we investigate whether the male-specific or female-specific vocabulary has a different impact on math performance (section 7.3).

7.1 Teacher Grading Bias

One first way through which teachers could encourage females and increase their performance is by overgrading them relative to their male peers. This teacher grading bias in favour of female students and its positive impact on school performance and enrollment choices has already been documented by Lavy and Sand (2018) and Terrier (2020). We use their methodology and estimate teachers' grading bias by taking the difference between the gender gap in math test scores at the Grade 12 continuous assessment and this gender gap at the math *baccalauréat* exam (see Appendix E for details). A negative (positive) grading bias is indicative of a grading bias in favor of females (males) at the continuous assessment. Consistent with what was previously found in the literature, we find that, on average, high school math teachers have a grading bias in favor of females (Table E7).

Turning to the correlation between the grading bias and our measure of teacher GDV, we see from Figure 15 Panel (a) that a one standard deviation increase in teacher GDV is associated with a -0.07 standard deviation decrease in the grading bias, significant at the one percent level. This correlation is moderate but suggests that teachers who differentiate their vocabulary more are also slightly more likely to encourage females through higher continuous assessment grades. However, controlling for the grading bias in the main specification does not affect the magnitude of the teacher GDV effect. Table 7 (Columns 1 and 3) reports the coefficients estimated on teacher GDV and on the teachers' grading bias for the main outcomes of interest.[22] Except for the percentile rank at the math *baccalauréat*, the teacher grading bias almost never significantly affects males' or females' outcomes and, in any case, its inclusion as a control does not change the coefficients on teacher GDV. We can therefore exclude teachers' grading bias as a mediator of the impact of teacher GDV.

7.2 Teacher Quality

We now investigate whether the positive effects resulting from exposure to a higher GDV teacher can go through teacher quality.

[22] See Table E8 in Appendix E for the teacher GDV coefficients on all the outcomes of interest when controlling for teacher grading bias or value-added.

We compute a measure of teacher value-added following the three steps described in the Chetty et al. (2014) paper. We first regress the percentile rank at the math *baccalauréat* exam on a set of students' baseline characteristics, variables accounting for students' past performance, and teachers' fixed effects. We predict residuals and use those students' residualized test scores 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. The coefficients obtained are finally used to predict teachers' value-added in year t. All the details of the estimation as well as the distribution of teacher value-added can be found in Appendix E.

We then check whether teachers' GDV is correlated with their quality. Panel (b) of Figure 15 suggests a small yet significant quadratic relationship, where teachers with GDV measures that are two standard deviations below or above average have a slightly lower value-added. We control for teacher quality in equation 6 and see that the impact of teacher GDV on percentile rank in math is slightly reduced with this additional control, going from +0.31 percentile rank to +0.25 percentile rank for males and from 0.37 percentile rank to 0.28 percentile rank for females (Table 7, columns 2 and 4.). Even though reduced, the impact of teacher GDV on students' math performance remains significant. This therefore suggests that our results are only partially channeled through teachers' quality.

7.3 Teacher's Type of Vocabulary Used

The last mechanism we explore is whether the type vocabulary used by teachers, i.e. the male or female-specific vocabulary, triggers different responses from students. For that purpose, we split the teacher GDV measure into two submeasures. For each teacher×class, we compute the share of correctly predicted females on the one hand (hereafter referred to as the GDV-females measure) and the share of correctly predicted males on the other hand (GDV-males measure). These two measures enable us to highlight different patterns in the vocabulary used by teachers:

1. GDV-females = GDV-males: a teacher for whom both measures are very close has an overall GDV that is equally due to the correct classification of males and females by the model;
2. GDV-females > GDV-males: in this case, the overall teacher GDV is predominantly explained by the fact that the teacher uses the female-specific vocabulary with its female students and therefore our model better predicts females over males for that teacher.
3. GDV-females < GDV-males: this is the reverse situation where the overall teacher GDV is predominantly explained by the use of the male-specific vocabulary with male students,

entailing a better prediction of males' gender over females'.

Panel (a) of Figure E4 displays the distributions of the overall *leave-one-out* GDV and of the *leave-one-out* GDVs computed on males and females separately.[23] This graph shows that males are more often correctly classified by our model given that, on average, 66 percent of males' gender is correctly predicted against 55 percent for females. Panel (b) plots the correlation between both GDV submeasures and further shows that teachers for whom one gender is often correctly predicted have a substantially lower proportion of correctly classified observations for the other gender. A one standard deviation increase in the teacher GDV-males measure is associated with a 0.7 standard deviation decrease in the GDV-females measure.

To measure the extent to which the positive effect on math performance at *baccalauréat* is due to teachers using mostly the vocabulary associated to females or that associated to males, we estimate Equation 6 replacing teacher GDV by both GDV-males and GDV-females submeasures. Table 8 reports the estimations results for the main outcomes, and shows that the positive effect on the percentile rank at the math *baccalauréat* documented for higher-GDV teachers is primarily driven by teachers with a higher GDV-females, i.e. teachers who are more inclined to use the vocabulary associated to females. A one standard deviation increase in teachers' GDV-females (GDV-males) is associated with a 0.49 (0.39) percentile rank increase at the math *baccalauréat* exam on average, controlling for the average GDV-males (GDV-females). The positive effect of being exposed to a higher GDV-females teacher is stronger for female students, for whom the percentile rank increases by 0.52 against 0.34 when exposed to a higher GDV-males teacher, while these estimates are respectively 0.48 and 0.45 for males. These results suggest that the vocabulary used by higher GDV-females teachers is more prone to raising female students' performance and, to a lower extent, that of males. These results should however be interpreted with caution, as differences between male and female students are not statistically significant.

The magnitude of coefficients on the probability of top-ranking a STEM program or on matriculation outcomes are too low to substantially change students' program ranking or matriculation in the following year.

8 Conclusion

Relying on text mining techniques, we explore the vocabulary used by Grade 12 math teachers to assess their male and female students' work and propose a synthetic measure of the gender

[23] Note that the density curve for the overall leave-one-out GDV is the same as the one displayed in Figure 5.

differentiation in teachers' vocabulary. This measure is designed to isolate feedback patterns that are independent from students' characteristics, and in particular from their ability level.

Using comprehensive administrative data on the universe of Grade 12 students' transcripts, we predict students' gender based on the feedback vocabulary, and compute the proportion of students whose gender is correctly predicted by our model separately for each teacher. This is our measure of teacher gender-differentiated vocabulary (GDV). We provide evidence that, on average, math teachers differentiate their vocabulary based on gender, and even more so when the teacher is a female. While this gendered-vocabulary exists in every core subjects, we find that teachers in science-related courses differentiate their vocabulary even more than teachers in humanities-related subjects. The qualitative analysis of the best gender predictors reveals that teachers insist more on positive managerial aspects and encourage the efforts provided by their female students, while equally performing males are both more criticized for their unruly behavior and praised for their intellectual skills. This gender differentiation is stronger as we move up in the teacher GDV distribution.

Exploiting the fact that, conditional on elective courses, the assignment of teachers within high school is quasi-random, we establish a causal link between teachers' GDV and students academic outcomes. Being exposed to a teacher with a one standard deviation higher GDV increases math performance at *baccalauréat* by 0.34 percentile rank on average, with slightly larger effects for female students. This effect is larger for students exposed to teachers with above-median GDV: relative to students exposed to a teacher from the bottom 10 percent of the GDV distribution, those exposed to above-median teachers see their percentile rank increase by up to 2 for females. A similar trend is observed for males, but to a lower extent. Deciles' coefficients range between 0 and 1.5 and are always below that of their female peers. We do not find any significant impact of teacher GDV on students' rank-order lists nor on students' matriculation in the following year.

Finally, we explore potential mechanisms driving the effects on math performance. We rule out the possibility that our results are driven by other teacher gendered behavior such as grading bias, but find that our results are partially channeled through teacher quality. We further show that our results are driven by teachers using the vocabulary associated to female students, i.e. teachers underlining the positive behavior and efforts, compared to teachers using the vocabulary associated to male students.

Our paper is at the crossroad between several social sciences, which offers a range of perspectives for future research. One pending question is the extent to which our measure of GDV relates to other concepts used in sociology or in psychology, such as the fixed *versus* growth mindsets, or the Implicit Association Test measure. These concepts could help understanding why insisting more on positive behavior and effort aspects induces better performance compared to a more competence-oriented vocabulary.

References

Alan, S., S. Ertac, and I. Mumcu, "Gender Stereotypes in the Classroom and Effects on Achievement," *Review of Economics and Statistics*, 2018, *100* (5), 876–890.

Beilock, S., E. Gunderson, G. Ramirez, and S. Levine, "Female Teachers' Math Anxiety Affects Girls' Math Achievement," *Proceedings of the National Academy of Science*, 2010, *107* (5), 1860–1863.

Bian, L., S. Leslie, and A. Cimpian, "Gender Stereotypes About Intellectual Ability Emerge Early and Influence Children's Interests," *Science*, 2017, *355*, 389–391.

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

Breda, T., E. Jouini, and C. Napp, "Societal Inequalities Amplify Gender Gaps in Math," *Science*, 2018, *359* (6381).

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

Cheng, A., , and Zamarro G. Kopotic K., "Can Parents' Growth Mindset and Role Modelling Address STEM Gender Gaps?," *Education Reform Faculty and Graduate Students Publications*, 2017.

Cheryan, S. and V. Plaut, "Explaining Underrepresentation: A Theory of Precluded Interest," *Sex Roles*, 2010, *63*, 475–488.

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.

Copur-Gencturk, Y., I. Thacker, and D. Quinn, "K-8 Teachers' Overall and Gender-Specific Beliefs About Mathematical Aptitude," *International Journal of Science and Mathematics Education*, 2020.

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.

Demoulin, H. and C. Daniel, "Bulletins scolaires et orientation au prisme du genre," *L'orientation scolaire et professionnelle*, 2013, *42.*

Dweck, C., W. Davidson, S. Nelson, and B. Enna, "Sex Differences in Learned Helplessness: II. The Contingencies of Evaluative Feedback in the Classroom and III. An Experimental Analysis," *Developmental Psychology*, 1978, *14* (3), 268–276.

Eble, Alex and Feng Hu, "How Important are Beliefs about Gender Differences in Math Ability? Transmission across Generations and Impacts on Child Outcomes," *EdWorkingPaper*, 2019, *19-67.*

_ and _ , "Stereotypes, Role Models, and the Formation of Beliefs," *CDEP-CGEG Working Paper No. 43*, 2019.

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

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

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.

Guiso, L., F. Monte, P. Sapienza, and L. Zingales, "Culture, Gender, and Math," *Science*, 2008, *320.*

Gunderson, E., G. Ramirez, S. Levine, and S. Beilock, "The Role of Parents and Teachers in the Development of Gender-Related Math Attitudes," *Sex Roles*, 2012, *66*, 153–166.

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

Keller, C., "Effect of Teachers' Stereotyping on Students' Stereotyping of Mathematics as a Male Domain," *The Journal of Social Psychology*, 2001, *141* (2), 165–173.

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

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.

Leslie, S., A. Cimpian, M. Meyer, and E. Freeland, "Expectations of Brilliance Underlie Gender Distribution Across Academic Discplines," *Science*, 2015, *347* (6219).

Lim, Jaegeum and Jonathan Meer, "The Impact of Teacher-Student Gender Matches: Random Assignment Evidence from South Korea," *Journal of Human Resources*, 2017, pp. 1215–7585.

Lindberg, S., J. Hyde, J. Peterson, and M. Linn, "New Trends in Gender and Mathematics Performance: A Meta-Analysis," *Psychological Bulletin*, 2010, *136* (6), 1123–1136.

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.

Miller, D., A. Eagly, and M. Linn, "Women's Representation in Science Predicts National Gender-Science Stereotypes: Evidence From 66 Nations," *Journal of Educational Psychology*, 2014, *107*, 631–644.

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).

Nosek, Brian A., Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, Selin Kesebir, Norbert Maliszewski, Félix Neto, Eero Olli, Jaihyun Park, Konrad Schnabel, Kimihiro Shiomura, Bogdan Tudor Tulbure, Reinout W. Wiers, Mónika Somogyi, Nazar Akrami, Bo Ekehammar, Michelangelo Vianello, Mahzarin R. Banaji, and Anthony G. Greenwald, "National Differences in Gender–Science Stereotypes Predict National Sex Differences in Science and Math Achievement," 2009, *106* (26), 10593–10597.

Oguzoglu, U. and S. Ozbeklik, "Like Father, Like Daughter (Unless There Is a Son): Sibling Sex Composition and Women's STEM Major Choice in College," *IZA Discussion Papers*, 2016, *10052.*

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

SkoǨajić, M., Radosavljević J., M. Okičić, I. Janković, and I. Žeželj, "Boys Just Don't! Gender-Stereotyping and Sacntionning of Counterstereotypical Behavior in Preschoolers," *Sex Roles*, 2020, *82*, 163–172.

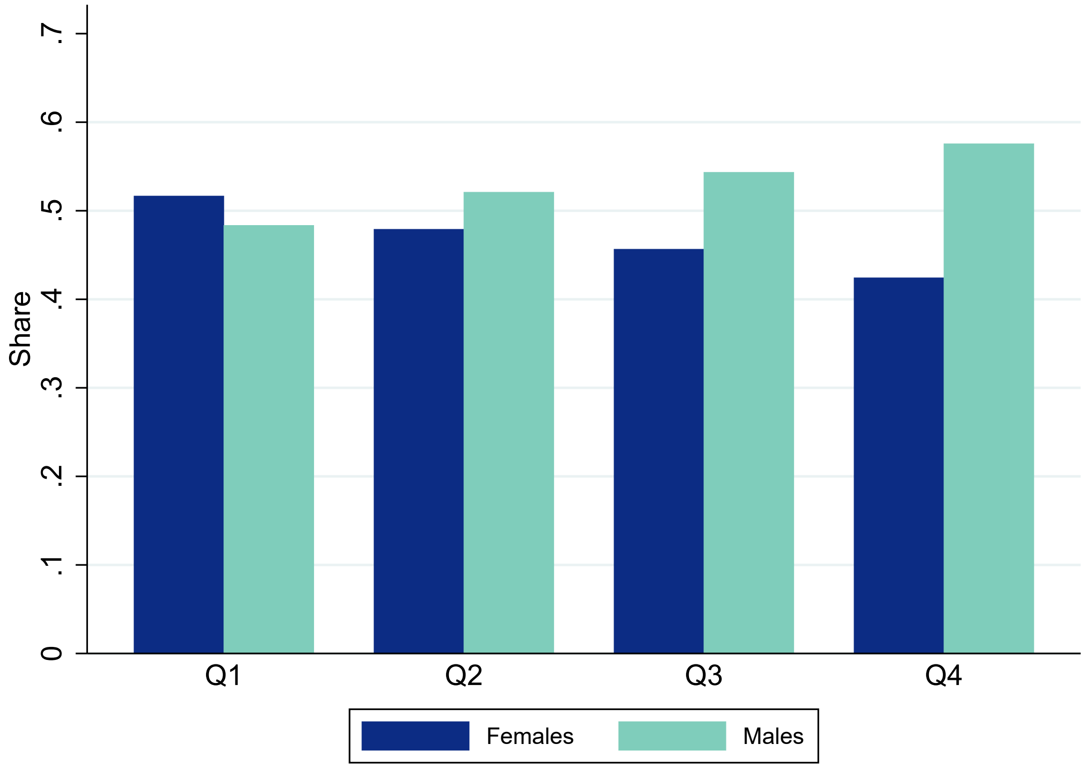
Stepner, M., "VAM: Stata Module to Compute Teacher Value-Added Measures," *Statistical Software Components*, 2013, *S457711.*

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.

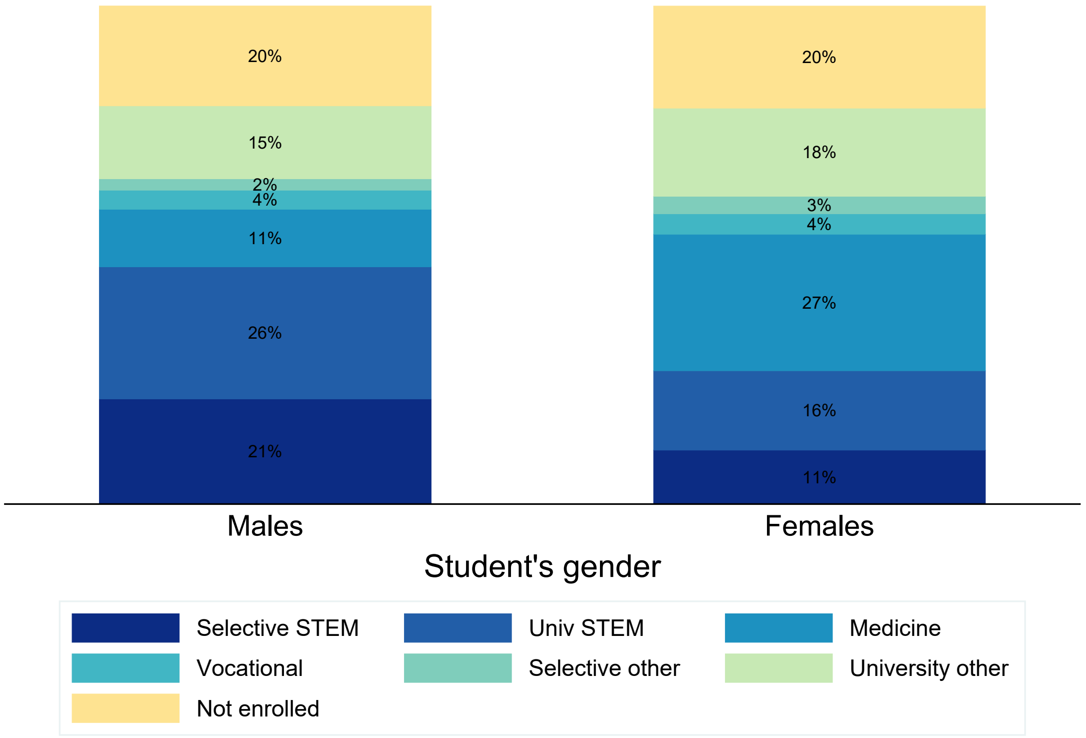
Wu, X., C. Li, Y. Zhu, and Y. Miao, "Short Text Topic Modeling with Topic Distribution Quantization and Negative Sampling Decoder," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Figure 1 – Share of Male and Female Students by Quartile of Percentile Rank at the DNB Math Exam



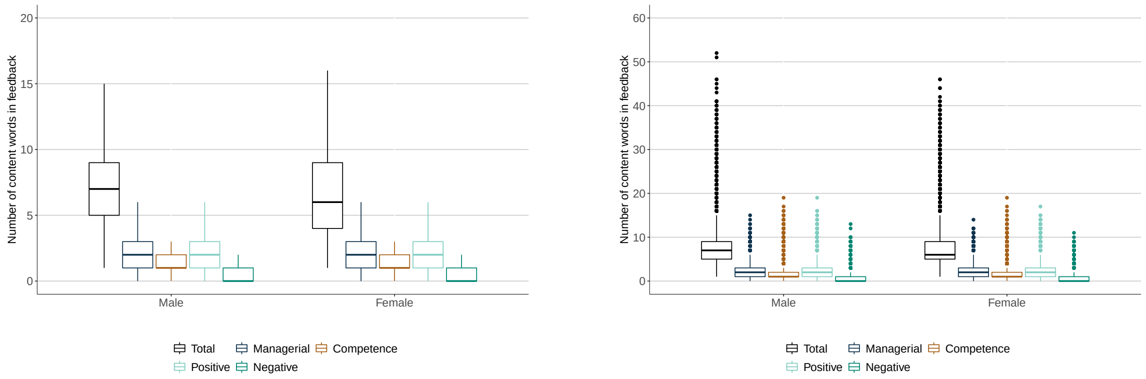
Notes: This graph shows the share of Grade 12 science major female and male students at each quartile of the math DNB percentile rank distribution, based on administrative data from the French Ministry of higher education.

Figure 2 – Enrollment in Higher Education by Gender after Grade 12 Science Track



Notes: This graph breaks down Grade 12 science major female and male students' matriculation choices into the main types of higher education programs, based on administrative data from the French Ministry of Education.

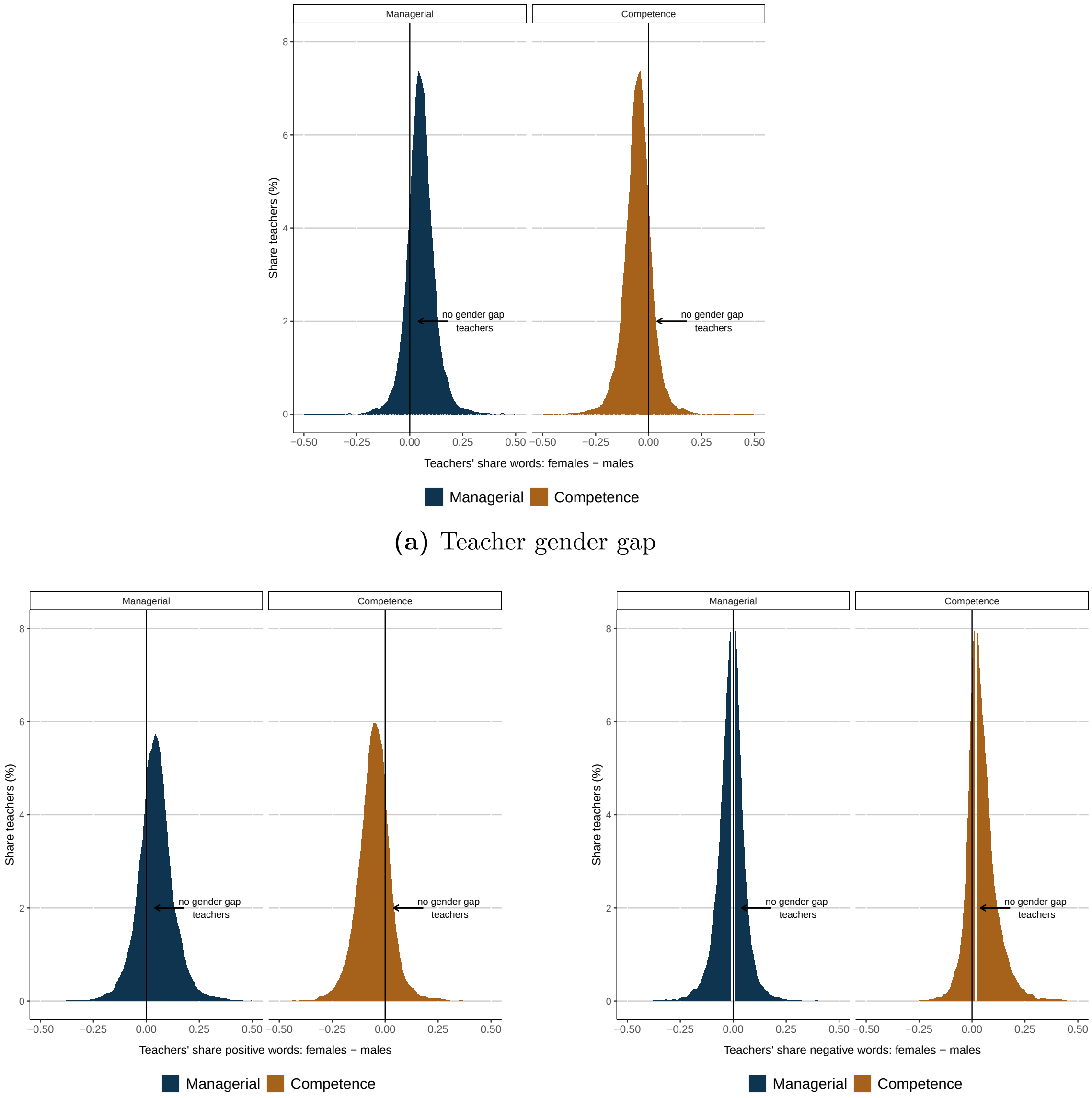
Figure 3 – Math Feedback - Distribution of Word Counts



(a) Summary statistics **(b)** Summary statistics and outliers values

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. Values outside of this range can be considered as outlier values, and Panel (b) displays each of these outlier values separately with a dot.

Figure 4 – Distribution of Teachers' Gender Gaps in Feedback Type and Positiveness - Females *minus* Males



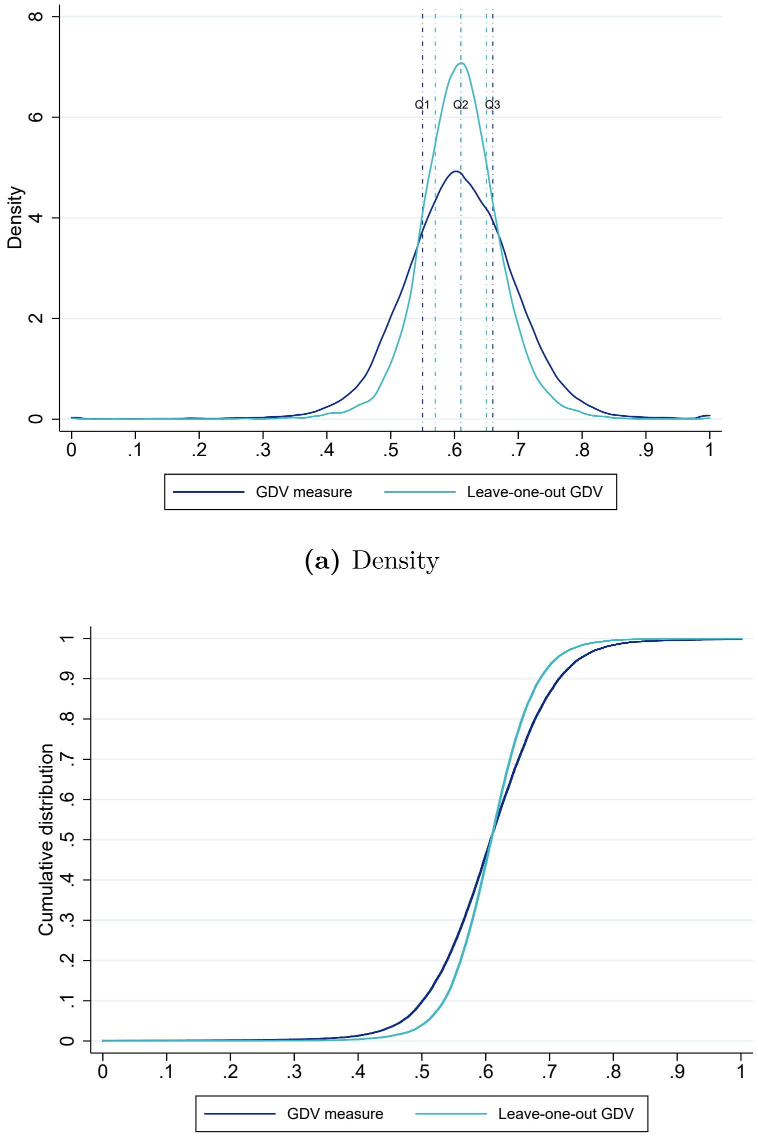
(a) Teacher gender gap

(b) Teacher gender gap in positive words

(c) Teacher gender gap in negative words

Notes: This figure shows the distributions of Grade 12 math teachers' gender gaps in feedback type and positiveness, based on administrative data from the French Ministry of higher education. Panel (a) displays the distributions of teachers' gender gaps in the share of Managerial and Competence words in the feedback given to their students. Panels (b) and (c) further detail the gender gaps distributions in the shares of positive and negative words, for each feedback type separately.

Figure 5 – Distribution of Math Teachers GDV and Leave-one-out GDV

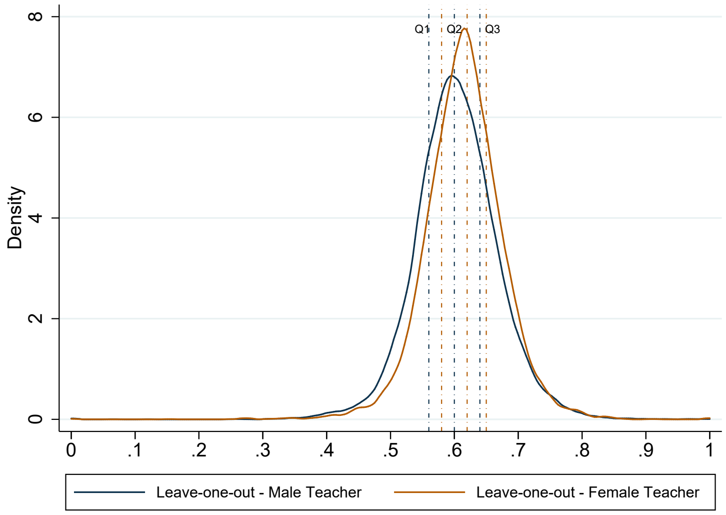


(a) Density

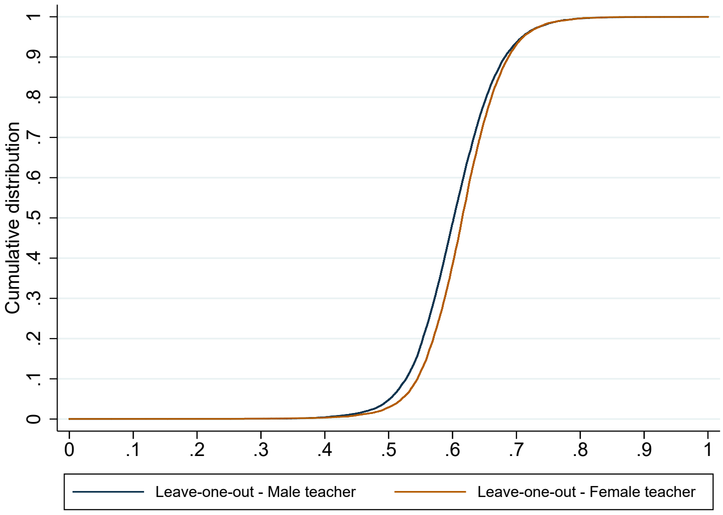
(b) Cumulative distribution

Notes: This figure shows the distributions of math teachers' GDV and leave-one-out GDV measures, 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 6 – Distribution of Math Teachers' Leave-one-out GDV – By Teacher Gender



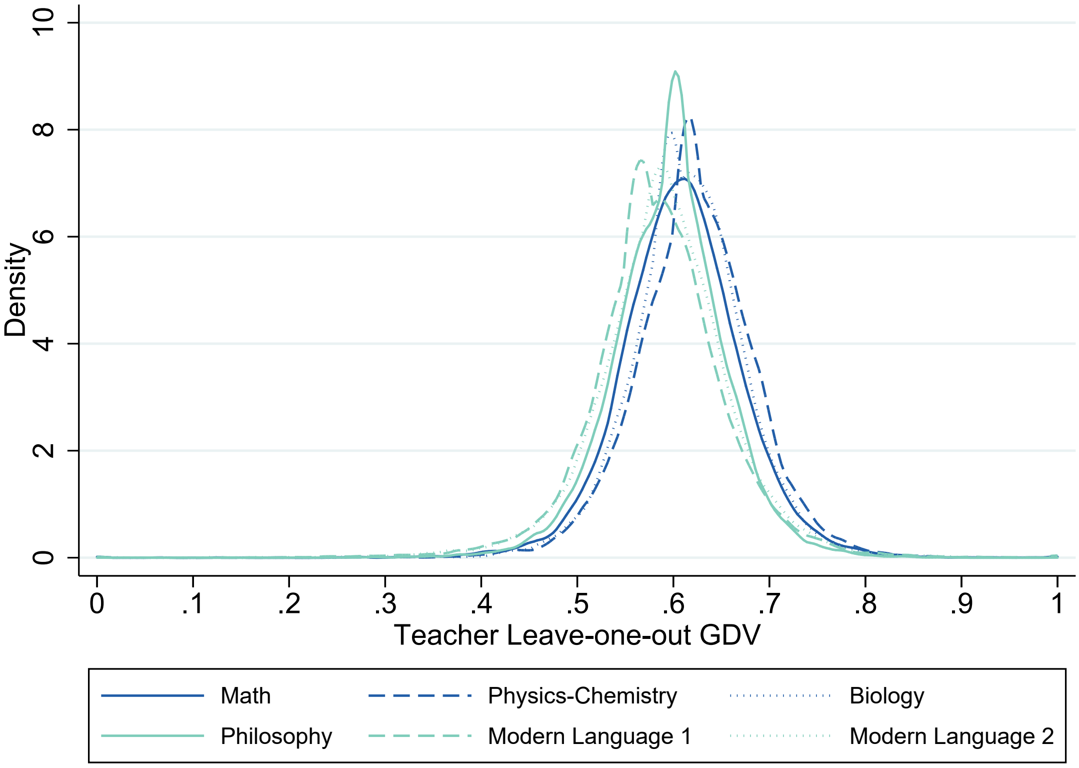
(a) Density



(b) Cumulative distribution

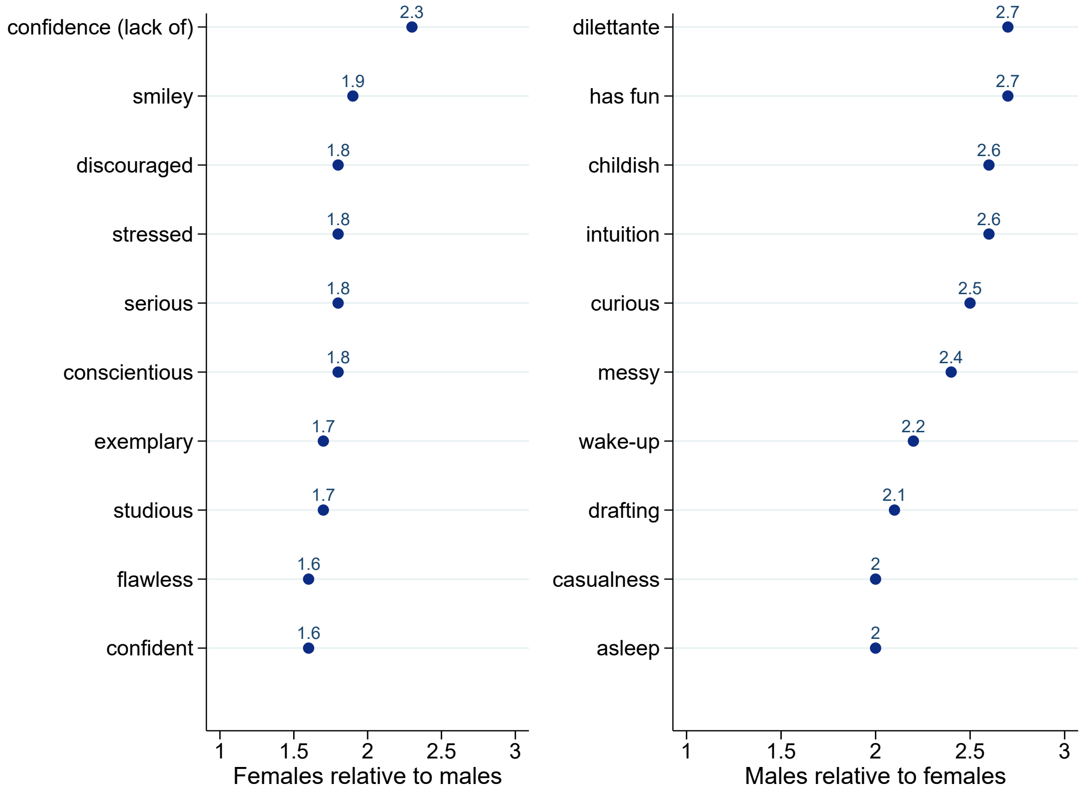
Notes: This figure shows the distributions of female and male math teachers' leave-one-out GDV measure, based on administrative data from the French Ministry of Education. The sample consists of Grade 12 math teachers teaching in high school × elective × year cells containing more than one teacher.

Figure 7 – Distribution of Teachers' Leave-one-out GDV – By Core Subjects



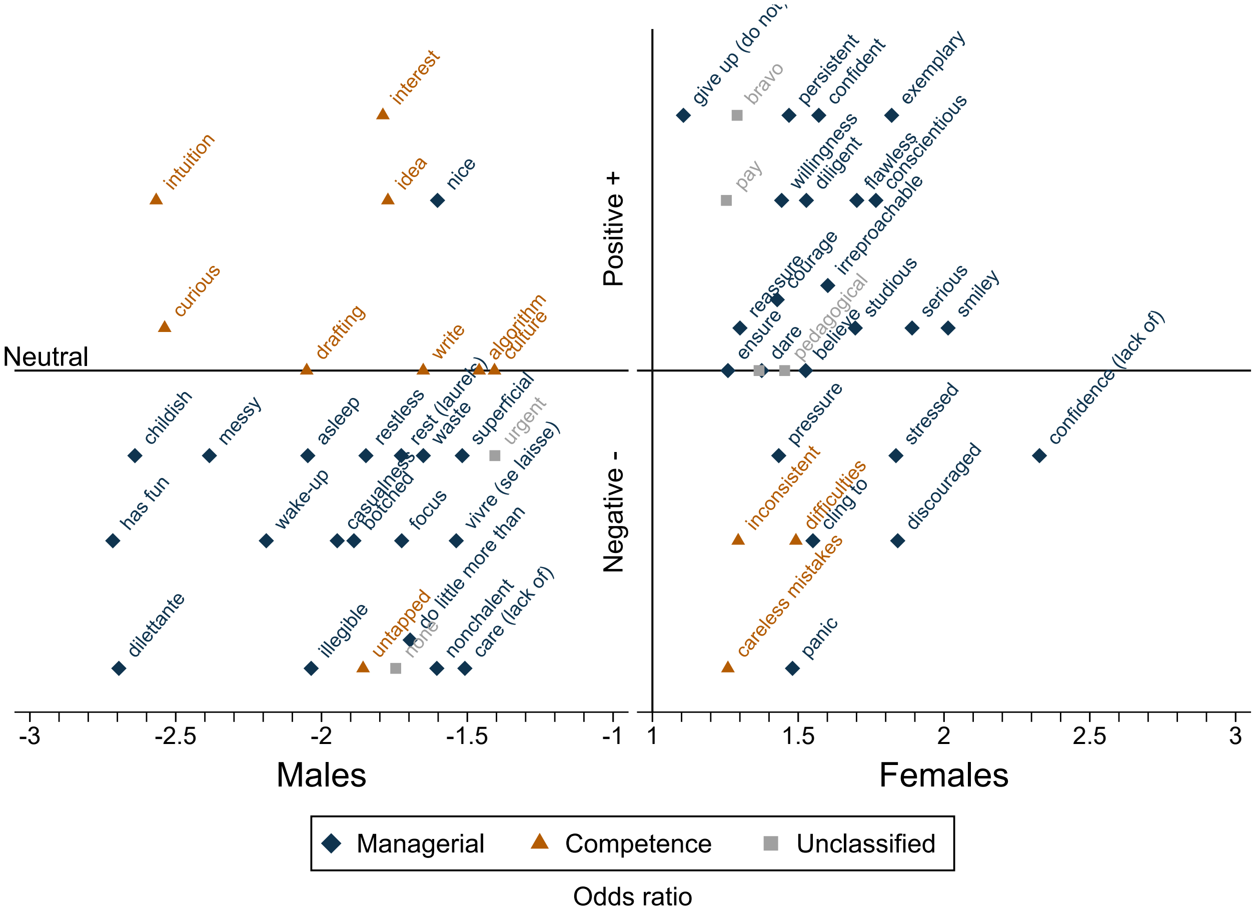
Notes: This figure shows the distributions of the math, physics, biology, philosophy and foreign language teachers' leave-one-out GDV measure, based on administrative data from the French Ministry of Education. The sample consists of Grade 12 teachers teaching in high school × elective × year cells containing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

Figure 8 – Odds Ratios of the Top 10 Gender Predictors



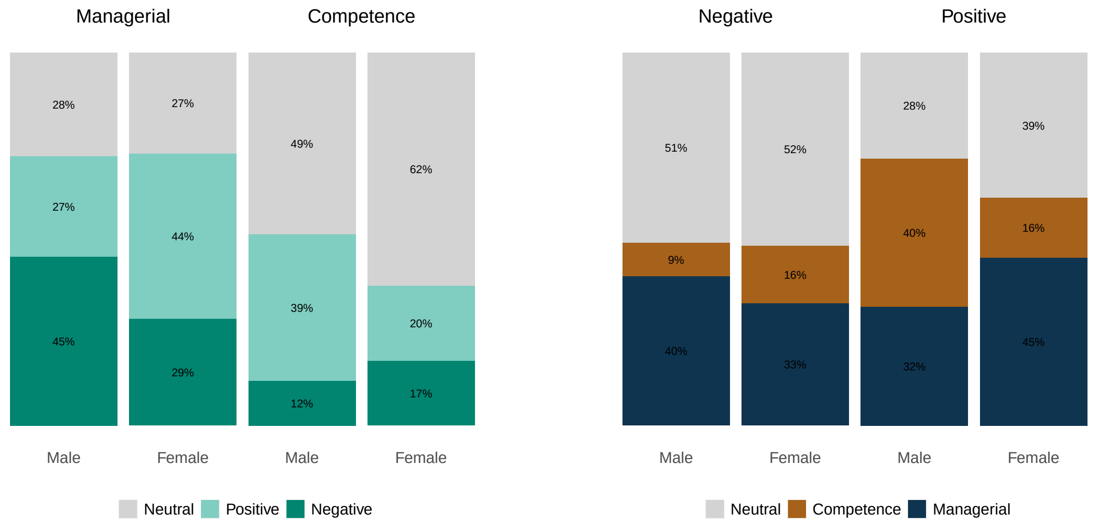
Notes: This figure shows the odds ratios obtained for the top 10 female and male predictors of the model described by Equation 1 estimated using the vocabulary appearing in math teachers' feedback. The estimation is realised on the universe of French Grade 12 science major students.

Figure 9 – 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.

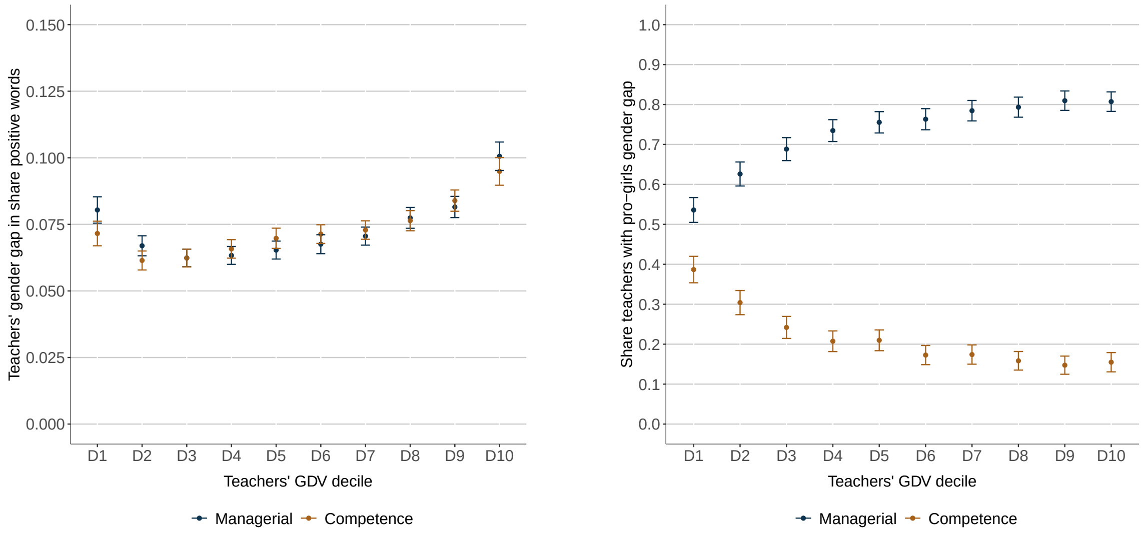
Figure 10 – Gender Predictors' Type and Positiveness



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

Notes: This figure shows the proportions of managerial and competence-related gender predictors conditional on positiveness (Panel a), and that of positive and negative gender predictors conditional on feedback type (Panel b).

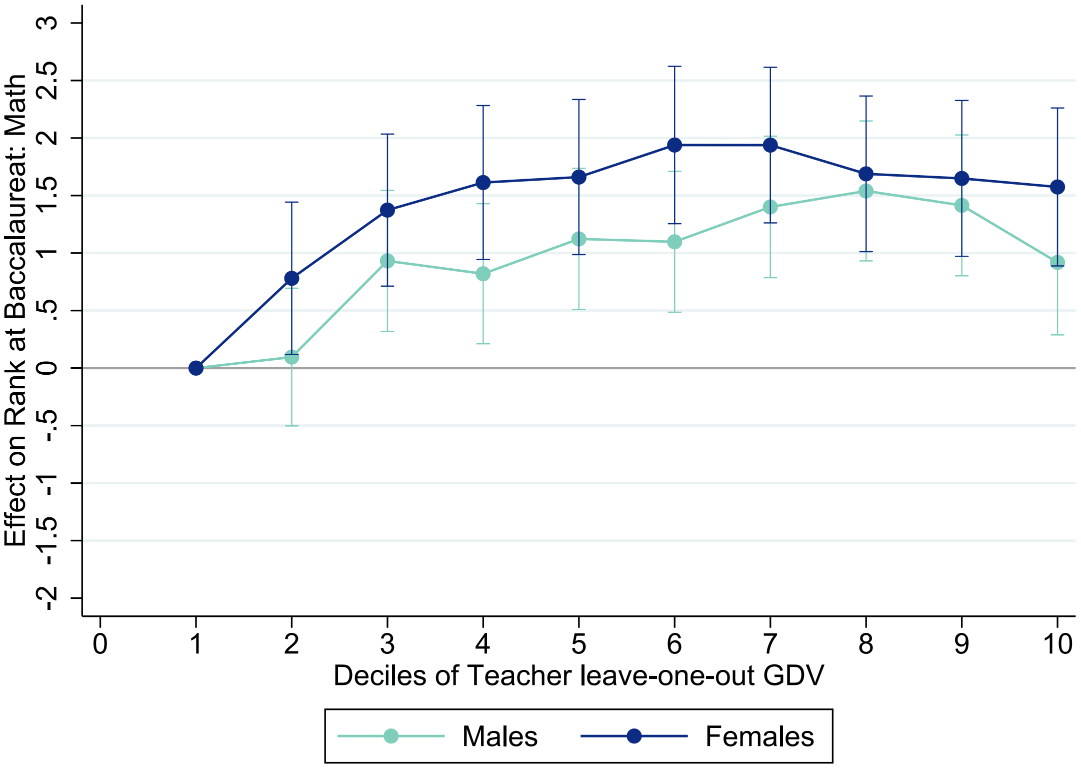
Figure 11 – Teachers' Gender Gap in the Share of Positive Words in Favour of Females by Deciles of GDV - In Absolute Value and Percentage



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

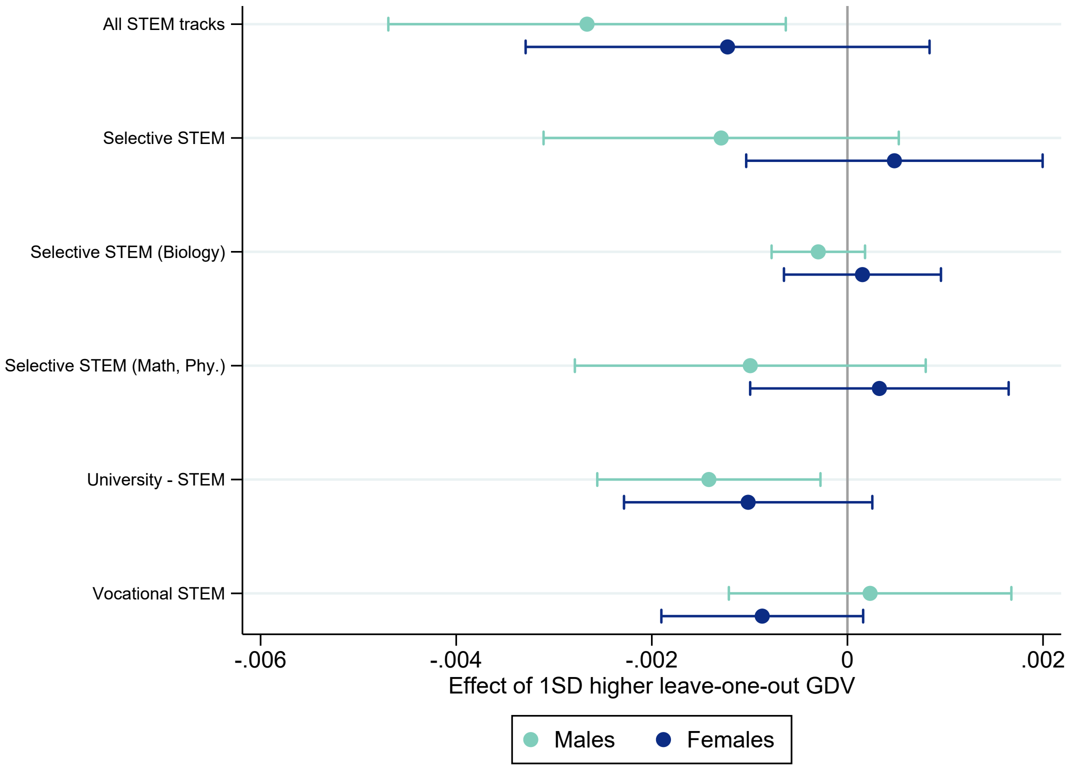
Notes: For each GDV decile, Panel (a) displays the average absolute value of Grade 12 teachers' gender gaps in the share of positive words appearing in their feedbacks, separately for competence vs. managerial related words. The GDV deciles are computed based on the teacher-level average of GDV measures. Panel (b) displays the share of teachers for whom the gender gap is in favour of females students, by GDV decile. The average values per decile are computed on the universe of math Grade 12 teachers for whom at least one GDV measure was estimated.

Figure 12 – Effect of Math Teacher GDV on Math Performance at *baccalauréat* - By GDV Deciles



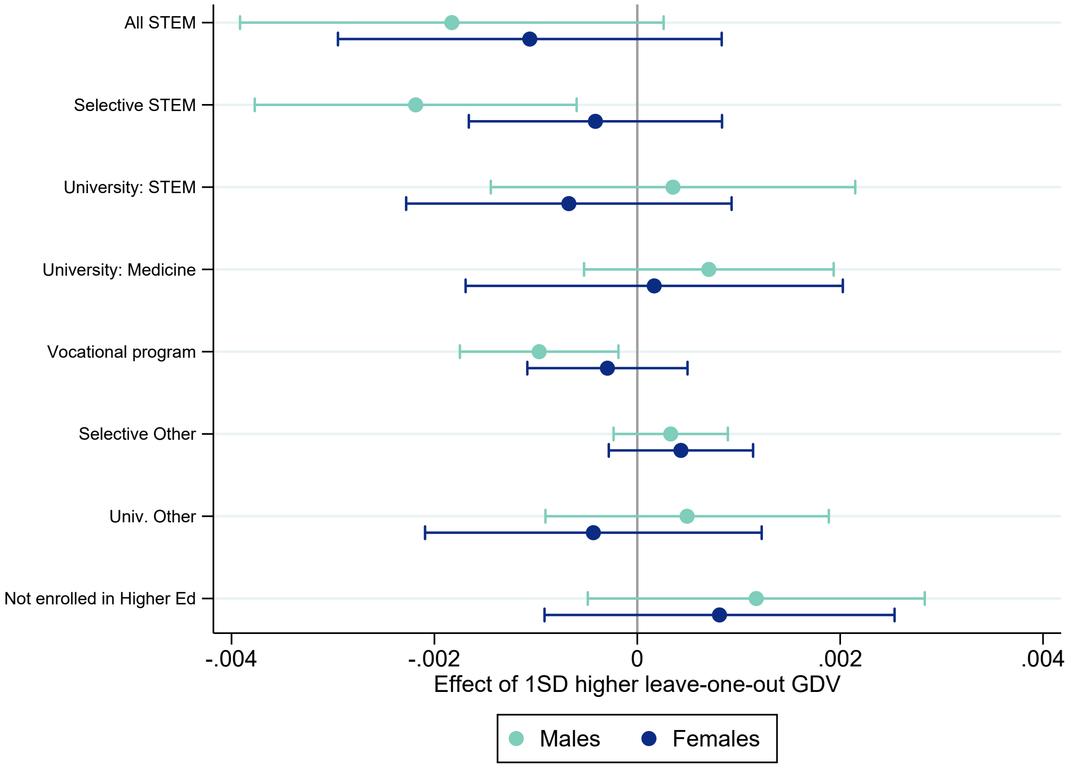
Notes: The results are calculated with administrative data from the French Ministry of higher education on French Grade 12 science major students. The figure reports the results of the regression of students' percentile rank at the math *baccalauréat* exam on a set of teacher leave-one-out GDV decile 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 on Grade 12 science major students for whom the high school × elective × year cell contains more than one math teacher.

Figure 13 – Effect of Math Teacher GDV on STEM Programs First Choice



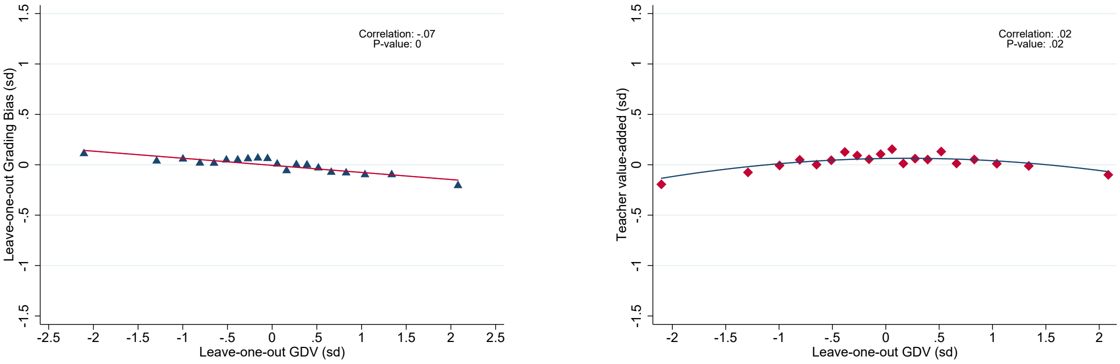
Notes: The results are calculated with administrative data from the French Ministry of higher education on French Grade 12 science major students. The figure shows the effect on female and male students' probability to top-rank STEM programs of being assigned a teacher with a one standard deviation higher leave-one-out GDV, along with its 95% confidence interval. The coefficients are estimated on students for whom the high school × elective × year cell contains more than one math teacher.

Figure 14 – Effect of Math Teacher GDV on Matriculation in the Following Year



Notes: The results are calculated with administrative data from the French Ministry of Education on French Grade 12 science major students. The figure shows the effect on female and male students' probability of matriculation in the following year of being assigned a teacher with a one standard deviation higher leave-one-out GDV. The coefficients are estimated on students for whom the high school × elective × year cell contains more than one math teacher.

Figure 15 – Correlation Between Teacher GDV, Grading Bias and Teacher Quality



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

Notes: The results are calculated with administrative data from the French Ministry of higher education on French Grade 12 science major students. 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 GDV. 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 GDV. The sample consists of all Grade 12 math teachers for whom a leave-one out GDV measure, a leave-one-out grading bias measure and a value-added measure could be estimated.

Table 1 – Number of Grade 12 Students and Sample Restrictions

	2012	**2013**	**2014**	**2015**	**2016**	**2017**
Total nb. of G12 students	174,996	179,625	183,693	190,980	198,573	203,262
Nb. of obs with missing transcript	90,299	79,226	54,248	42,502	34,064	28,599
% high school entirely missing:	95.7	92.7	91.0	85.2	78.2	67.8
High school < 2 classes	3,634	4,716	6,443	6,836	7,516	7,680
Teachers < 2 classes	14,165	5,744	5,930	5,864	8,860	32,089
Obs. in the analytical sample	**66,898**	**89,939**	**117,072**	**135,778**	**148,133**	**134,894**
(in %)	(38.2)	(50.1)	(63.7)	(71.1)	(74.6)	(66.4)

Notes: This table reports the number of Grade 12 students appearing each year in the APB database. 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 – Students' Summary Statistics

	All	**Males**	**Females**
Demographics			
Female student (N= 691,093)	0.47		
Age (years) (N= 691,093)	18.09	18.12	18.06
Free lunch student (N= 691,059)	0.13	0.12	0.14
High SES (N= 691,093)	0.43	0.44	0.41
Medium-high SES (N= 691,093)	0.16	0.16	0.16
Medium-low SES (N= 691,093)	0.24	0.24	0.25
Low SES (N= 691,093)	0.17	0.16	0.18
Education			
Rank at DNB: math (N= 654,958)	50.28	52.18	48.13
Rank at DNB: French (N= 654,927)	50.33	44.69	56.73
Rank at *baccalauréat*: French (written) (N= 659,291)	50.00	45.01	55.62
Rank at *baccalauréat*: French (oral) (N= 659,254)	49.78	45.69	54.39
Maths elective (N= 622,903)	0.23	0.27	0.19
Physics-Chemistry elective (N= 622,903)	0.26	0.27	0.25
Earth & Life Science elective (N= 622,903)	0.37	0.26	0.50
Engineering & Info elective (N= 622,903)	0.13	0.20	0.06
Nb of classmates (N= 691,093)	30.83	30.68	31.01
Nb. of observations	691,093	368,922	322,171

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

Table 3 – Teachers' Summary Statistics

	Mean	S.D
Share of head teacher at least once (N= 6,754)	0.53	0.50
Male math teacher (N= 6,721)	0.58	0.49
Number of teacher observations (N= 6,772)	3.70	1.64
Average number of classes per year (N= 6,772)	1.09	0.26
Average number of students per class (N= 6,772)	28.02	5.22
Average feedback length (N= 6,759)	7.40	2.51
Nb. of teachers	6,772	

Notes: This table shows descriptive statistics for the Grade 12 math teachers in the analytical sample. 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: Teachers' Leave-One-Out GDV with Students' Baseline Characteristics

	Dep. var: Teacher's leave-one-out GDV (sd)		
	Coeff.	**S.e**	**p-value**
Female student	−0.0028	0.0028	0.3184
Age (years)	−0.0043*	0.0026	0.0994
Free lunch student	0.0041	0.0037	0.2719
Foreign student	0.0055	0.0087	0.5269
High SES	−0.3163	0.3123	0.3112
Medium-high SES	−0.3155	0.3124	0.3125
Medium-low SES	−0.3193	0.3122	0.3063
Low SES	−0.3174	0.3124	0.3096
Rank at DNB: math	−0.0001	0.0000	0.2922
Rank at DNB: French	0.0000	0.0001	0.5023
Rank at *baccalauréat*: French (written)	0.0001	0.0001	0.2298
Rank at *baccalauréat*: French (oral)	0.0001	0.0001	0.2421
High school, elective, year FE	Yes		
F-stat (p-value)	1.12	(0.334)	
Nb. of observations	573,595		

Notes: This table reports the estimation results of the teachers' standardized leave-one-out GDV measure, defined at the class-level, regressed on the students' socio-economic characteristics and baseline academic performance. The regression includes high school, year and elective course fixed-effects. Standard errors are clustered at the teacher level and are reported in the second column. The F-statistics test for the joint significance of regressors. ***: 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	**Nb. of significant**	**Share sig. at**	
	p-values	**p-values at 5%**	**5%**	**1%**
	(1)	(2)	(3)	(4)
Female	21,935	2,483	11.32	3.37
Age (years)	19,800	1,591	8.04	2.60
Free-lunch	19,170	973	5.08	1.29
Foreign student	8,094	328	4.05	1.36
High SES	22,151	1,489	6.72	1.44
Medium-high SES	20,847	929	4.46	0.78
Medium-low SES	21,928	1,156	5.27	0.93
Low SES	20,394	1,127	5.53	1.15
Rank at DNB: math	22,485	1,377	6.12	1.21
Rank at DNB: French	22,489	1,534	6.82	1.34
Rank at *baccalauréat*: French (written)	22,482	1,640	7.29	1.58
Rank at *baccalauréat*: French (oral)	22,475	1,613	7.18	1.41

Notes: This table reports the results of the Pearson Chi-square tests of independance performed on the 27,688 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% and 1% respectively.

Table 6 – Impact of Teacher GDV on Academic Performance, Higher Education Choices and Enrollment

	All	**Males**	**Females**
	(1)	(2)	(3)
Academic performance			
Rank at *baccalauréat*: math	0.3351***	0.3113***	0.3735***
	(0.0668)	(0.0747)	(0.0811)
Rank at *baccalauréat*: philo	−0.0235	−0.0381	−0.0093
	(0.0646)	(0.0746)	(0.0786)
Type of STEM programs ranked first in the ROL			
All STEM tracks	−0.0019**	−0.0027**	−0.0012
	(0.0008)	(0.0010)	(0.0011)
Selective STEM	−0.0005	−0.0013	0.0005
	(0.0007)	(0.0009)	(0.0008)
among which: biology	−0.0001	−0.0003	0.0002
	(0.0002)	(0.0002)	(0.0004)
among which: math, physics	−0.0004	−0.0010	0.0003
	(0.0007)	(0.0009)	(0.0007)
University - STEM	−0.0012***	−0.0014**	−0.0010
	(0.0004)	(0.0006)	(0.0006)
Vocational STEM	−0.0003	0.0002	−0.0009*
	(0.0005)	(0.0007)	(0.0005)
Matriculation in the following year			
All STEM	−0.0014	−0.0018*	−0.0011
	(0.0008)	(0.0011)	(0.0010)
Selective STEM	−0.0013**	−0.0022***	−0.0004
	(0.0006)	(0.0008)	(0.0006)
University STEM	−0.0001	0.0004	−0.0007
	(0.0007)	(0.0009)	(0.0008)
University Medicine	0.0004	0.0007	0.0002
	(0.0006)	(0.0006)	(0.0009)
Vocational program	−0.0007**	−0.0010**	−0.0003
	(0.0003)	(0.0004)	(0.0004)
Selective Other	0.0004*	0.0003	0.0004
	(0.0002)	(0.0003)	(0.0004)
University Other	0.0001	0.0005	−0.0004
	(0.0006)	(0.0007)	(0.0008)
Not enrolled in Higher Ed	0.0010	0.0012	0.0008
	(0.0007)	(0.0008)	(0.0009)
Nb. observations	649,105	345,201	303,904

Notes: Each row reports the coefficients of the standardized *leave-one-out* teacher GDV obtained from the estimation of Equation 6 for the different outcomes listed on the first column. It is estimated on all the sample and separately for Grade 12 males and females. The regression includes high school, year and elective course 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 – Impact of Teacher GDV - Mechanisms

	Males		**Females**	
	Grading bias	Value added	Grading bias	Value added
	(1)	(2)	(3)	(4)
Academic performance				
Rank at Baccalaureat: Math	0.3714***	0.2486***	0.3967***	0.2825***
	(0.0770)	(0.0684)	(0.0851)	(0.0727)
Coeff. on grading bias or VA	0.5593***	2.8396***	−0.1437*	3.6506***
	(0.0788)	(0.1420)	(0.0861)	(0.1503)
Type of STEM programs ranked first in the ROL				
All STEM tracks	−0.0029***	−0.0027**	−0.0015	−0.0015
	(0.0011)	(0.0011)	(0.0011)	(0.0010)
Coeff. on grading bias or VA	−0.0010	0.0014	−0.0002	−0.0013
	(0.0011)	(0.0018)	(0.0011)	(0.0016)
Selective STEM	−0.0015	−0.0014	0.0003	0.0003
	(0.0010)	(0.0010)	(0.0008)	(0.0008)
Coeff. on grading bias or VA	−0.0004	0.0036**	0.0001	0.0004
	(0.0010)	(0.0017)	(0.0008)	(0.0013)
University - STEM	−0.0015***	−0.0014**	−0.0010	−0.0010
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Coeff. on grading bias or VA	−0.0003	−0.0021**	−0.0008	−0.0020*
	(0.0006)	(0.0010)	(0.0007)	(0.0010)
Matriculation in the following year				
All STEM	−0.0022**	−0.0023**	−0.0014	−0.0014
	(0.0011)	(0.0011)	(0.0010)	(0.0010)
Coeff. on grading bias or VA	−0.0010	0.0051***	−0.0003	0.0010
	(0.0011)	(0.0018)	(0.0010)	(0.0015)
Selective STEM	−0.0025***	−0.0025***	−0.0004	−0.0005
	(0.0008)	(0.0008)	(0.0007)	(0.0007)
Coeff. on grading bias or VA	0.0015*	0.0038***	0.0002	0.0003
	(0.0009)	(0.0015)	(0.0007)	(0.0011)
University: STEM	0.0002	0.0002	−0.0010	−0.0010
	(0.0010)	(0.0010)	(0.0008)	(0.0008)
Coeff. on grading bias or VA	−0.0023**	0.0015	−0.0005	0.0007
	(0.0010)	(0.0015)	(0.0009)	(0.0013)
Nb. observations	340,729	342,380	300,020	301,409

Notes:

Table 8 – Impact of Teacher GDV-Females and GDV-Males - Mechanisms

	All		**Males**		**Females**	
	GDV Males	GDV Females	GDV Males	GDV Females	GDV Males	GDV Females
	(1)	(2)	(3)	(4)	(5)	(6)
Academic performance						
Rank at *baccalauréat*: math	0.3921***	0.4905***	0.4497***	0.4828***	0.3358***	0.5232***
	(0.0854)	(0.0904)	(0.0953)	(0.0986)	(0.1040)	(0.1110)
Type of STEM programs ranked first in the ROL						
All STEM tracks	−0.0017	−0.0022**	−0.0028**	−0.0029**	−0.0007	−0.0014
	(0.0011)	(0.0011)	(0.0014)	(0.0014)	(0.0013)	(0.0013)
Selective STEM	−0.0010	0.0004	−0.0019	−0.0005	0.0002	0.0016
	(0.0009)	(0.0009)	(0.0012)	(0.0013)	(0.0010)	(0.0010)
University STEM	−0.0003	−0.0020***	−0.0006	−0.0025***	−0.0002	−0.0016*
	(0.0006)	(0.0006)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Matriculation in the following year						
All STEM	−0.0017	−0.0014	−0.0024*	−0.0017	−0.0011	−0.0013
	(0.0011)	(0.0011)	(0.0014)	(0.0014)	(0.0012)	(0.0012)
Selective STEM	−0.0021***	−0.0001	−0.0035***	−0.0008	−0.0005	0.0005
	(0.0008)	(0.0008)	(0.0011)	(0.0011)	(0.0009)	(0.0008)
University STEM	0.0004	−0.0014	0.0012	−0.0010	−0.0006	−0.0019*
	(0.0009)	(0.0009)	(0.0012)	(0.0012)	(0.0010)	(0.0010)
Nb. observations	642,584	642,433	341,168	341,111	301,416	301,322

Notes: Each row reports the coefficient of the standardized *leave-one-out* teacher GDV-males and GDV-females measures obtained from the estimation of Equation 6 for the different outcomes listed on the first column. It is estimated on all the sample and separately for Grade 12 males and females. The regression includes high school, year and elective course 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

Teacher Gendered Feedback, Students' Math Performance and Enrollment Outcomes

Pauline Charousset, Marion Monnet

May 2021

List of Appendices

A Measuring Teachers' GDV: Details of the Estimation Procedure

This appendix provides the details on the practical implementation for the different steps of teachers' gender differentiated feedback (GDV) estimation procedure developped in Section 4.

A.1 Textual Data Preparation

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

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

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

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

Table A1 – From text to data: an illustration

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

Token classification. Building on the considerable literature in psychology and sociology related to the analysis of gendered feedback (Dweck et al., 1978; Morgan, 2001), we classify the gender predictors into one of the four following categories: positive (resp. negative) competence-related aspects and positive (resp. negative) managerial aspects. We classify as competence-related any word that either relates to math and the school environment (e.g. reasoning, exam, geometry) or to an intellectual skill (e.g. talented, potential). 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 labeled neutral or unclassified. The words classified as managerial are the adjectives used to describe the student's behavior in class (e.g. shy, exemplary) as well as the actions undertaken or the efforts provided by the student (e.g. involved, revise, fall behind). The classification of the 100 best predictors of each gender can be found in Tables A2 and A3.[A.1][A.2]

A.2 Predicting Student Gender and Measuring Teacher GDV

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

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

and our objective is to find the set of α coefficients that minimize the penalized log-likelihood function, where λ is the regularization parameter:

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

The $\hat{\alpha}$ estimates are then used to predict students' gender. The teacher's GDV measure is computed based on those predictions, and is defined as the proportion of students for whom the model correctly predicts their gender, separately for each teacher. In practice, we estimate a logistic-Lasso to determine the α coefficients. We detail below the practical implementation of the estimation.

Step 1: Undersampling. Before any estimation is done, we deal with the issue of gender imbalance using undersampling techniques. Because gender is correlated with math performance, the model is likely to perform better on classes with larger gender imbalances in terms of math performance. For that reason, for each class, we sample as many males and females from each quartile of prior math performance. We define quartiles of prior math performance as follows: within each Grade 12 class, we rank students according to the math grade obtained at the DNB exam and create quartiles. 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.3]

[A.1] Even though every token has been classified, we only show the top 100 predictors given that other predictors are not more frequently used for female or male students (their odds ratio is around 1) and cannot be classified in any of the four mentioned categories in the vast majority of cases.

[A.2] We also attempted to build these categories in a data-driven manner using bi-term topic models tailored for short texts, but these models performed poorly on our data. Our data is indeed quite specific in that texts are very short, with an average number of tokens equal to 7, the overall vocabulary is quite limited ($\simeq$ 1,600 words) with little variation in the topics used as they all relate to academic performance and behavior. We therefore faced the typical challenges inherent to such short texts: the generated topics gathered inconsistent words (*trivial topics*) and the different topics were highly similar with a lot of words in common (*repetitive topics*, see Wu et al. (2020) for a discussion on those issues.)

[A.3] We use the French grade obtained at the DNB exam instead of the math grade when we compute the teacher GDV for humanities related subjects, i.e. for philosophy and modern languages.

Step 2: Random selection of tokens. As shown in Table 3, the number of tokens used in feedback varies by teacher. As feedback length could influence the quality of the prediction, we randomly sample tokens for lengthy feedback, defined as the ones with an above-median length. For such feedback we randomly select six tokens, which is the median number.

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

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

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

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

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

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

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

The two GDV 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 GDV measures averaged over those 100 iterations.

Table A2 – Top 100 Predictors' Classification - Female

	Positive	**Negative**	**Neutral**
Competence-related	3 tokens: autonomous, master, quality	5 tokens: careless mistakes, difficulties, inconsistent, mistake, misunderstandings	10 tokens: appropriate, calculus, classical, guidelines, method, question, read, support, usual
Managerial	29 tokens: cling to, confident, conscientious, courage, deserve, determined, diligent, discrete, efficient, encourage, exemplary, flawless, give up (do not), impeccable, irreproachable, keep doing, pay, persevere, persistent, reassure, reward, serious, seriously, smiley, steady, studious, tenacious, voluntary, willingness	15 tokens: cling to, concern, confidence (lack of), discouraged, doubt, give up, hesitate, panic, pressure, shy, stressed, suffer, unassuming, worry	6 tokens: believe, dare, ensure, intervene, pursue
Unclassified	6 tokens: bravo, congratulations, fruit (of work), laudable, pay, positive	5 tokens: decrease, drop, fragile, mishap, too low	22 tokens: a lot, allow, also, benchmark, big, circumstances, complete, context, contribute, despite, health, help, illustrate, know, other, pedagogical, point, pupil, recover, regular

Notes: This table reports the classification of the 100 best female feedback predictors into one of the four following categories: positive (resp. negative) competence-related aspects and positive (resp. negative) managerial aspects. We classify as competence-related any word that either relates to math and the school environment (e.g. reasoning, exam, geometry) or to an intellectual skill (e.g. talented, potential). The words classified as managerial are the adjectives used to describe the student's behaviour in class (e.g. shy, exemplary) as well as the actions undertaken or the efforts provided by the student (e.g. involved, revise, fall behind). 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.

Table A3 – Top 100 Predictors' Classification - Male

	Positive	**Negative**	**Neutral**
Competence-related	15 tokens: ambition, aptitude, capability, capacities, curious, idea, interest, intuition, passion, potential, relevant, rigorous, rigourous, scientific, sharpness	3 tokens: mix up, slow, untapped	13 tokens: algorithm, argument, computing, culture, drafting, expression (oral/written), guidelines, homework, passage, word, write, writing, written
Managerial	5 tokens: consciousness, detailed, nice, reaction, rebound	28 tokens: asleep, botched, care (lack of), casualness, childish, dilettante, disorganized, do little more than, focus, has fun, illegible, immature, inexistant, messy, minimal, nonchalent, rest (laurels), restless, scattered, shake up, skim through, superficial, troublesome, unacceptable, vivre (se laisse), wake-up, waste	5 tokens: behave, exploit, intervene, jusitfy, work
Unclassified	3 tokens: best, easy, sufficient	8 tokens: excessive, insufficient, minimum, none, perfectible, shame, sufficient, urgent	21 tokens: advice, could, decide, double, expected, handed in, imposed, invite, mature, measure, obvious, outside, personal, put, radical, time, took, want

Notes: This table reports the classification of the 100 best male feedback predictors into one of the four following categories: positive (resp. negative) competence-related aspects and positive (resp. negative) managerial aspects. We classify as competence-related any word that either relates to math and the school environment (e.g. reasoning, exam, geometry) or to an intellectual skill (e.g. talented, potential). The words classified as managerial are the adjectives used to describe the student's behaviour in class (e.g. shy, exemplary) as well as the actions undertaken or the efforts provided by the student (e.g. involved, revise, fall behind). 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.

B Teacher Gender Composition - All Core Subjects

Table B4 – Share of Male Teachers by Core Subjects

Subject	**Share**	**N**	**% non-miss**
Math	0.58	7,124	0.93
Physics-Chemistry	0.57	7,760	0.93
Biology	0.37	6,694	0.92
Philosophy	0.62	7,420	0.95
Modern language 1	0.20	17,589	0.88
Modern language 2	0.19	22,613	0.83

Notes: This table reports the share of male teachers for each core subject taught to Grade 12 students in our sample. The second column gives the number of teachers and the third column gives the proportion of teachers for whom the sex was not missing.

C Assessing the Randomness of Missing Grade Transcripts

Table C5 – 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	−0.1167***	0.0362	0.0013
Age (years)	0.1875***	0.0141	0.0000
Free lunch student	−0.3854***	0.0485	0.0000
Foreign student	−0.0002	0.0871	0.9981
High SES	0.0110	0.0421	0.7938
Medium-high SES	−0.3664***	0.0691	0.0000
Medium-low SES	−0.1253**	0.0535	0.0191
Rank at DNB maths	0.0026	0.0021	0.2309
Rank at DNB math (females)	0.0011	0.0010	0.2599
Rank at DNB math (males)	−0.0039***	0.0013	0.0023
Nb. of observations	12,864		

Notes: This table reports the estimation results of dummies 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.

D Robustness Checks and Additional Results

D.1 Robustness Checks

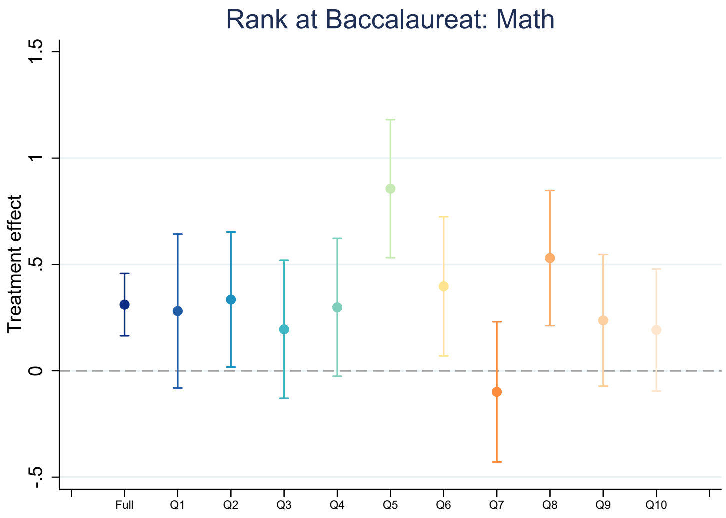
Table D6 – Impact of Teacher GDV - Robustness Checks

	Males			**Females**		
	Baseline X	*% of females*	*Average GDV*	*Baseline X*	*% of females*	*Average GDV*
	(1)	(2)	(3)	(4)	(5)	(6)
Academic performance						
Rank at *baccalauréat*: math	0.3541^{***}	0.3979^{***}	0.3985^{***}	0.4978^{***}	0.5027^{***}	0.5047^{***}
	(0.0690	(0.0744	(0.0743)	(0.0730)	(0.0810)	(0.0811)
Rank at *baccalauréat*: philo	−0.0797	−0.0233	−0.0263	−0.0691	−0.0432	−0.0455
	(0.0689	(0.0745	(0.0745)	(0.0694)	(0.0793)	(0.0793)
Type of STEM programs ranked first in the ROL						
All STEM tracks	-0.0025^{**}	-0.0025^{**}	-0.0023^{**}	−0.0015	−0.0013	−0.0013
	(0.0011	(0.0010	(0.0010)	(0.0011)	(0.0011)	(0.0011)
Selective STEM	-0.0016^{*}	−0.0013	−0.0013	0.0005	0.0004	0.0004
	(0.0009	(0.0009	(0.0009)	(0.0008)	(0.0008)	(0.0008)
among which: biology	-0.0004^{*}	−0.0004	−0.0004	0.0001	0.0000	0.0000
	(0.0003	(0.0002	(0.0002)	(0.0004)	(0.0004)	(0.0004)
among which: math, physics	−0.0012	−0.0010	−0.0009	0.0004	0.0004	0.0004
	(0.0009	(0.0009	(0.0009)	(0.0007)	(0.0007)	(0.0007)
University STEM	-0.0010^{*}	-0.0012^{*}	-0.0012^{*}	-0.0014^{**}	−0.0011	−0.0011
	(0.0006	(0.0006	(0.0006)	(0.0007)	(0.0007)	(0.0007)
Vocational STEM	0.0003	0.0002	0.0002	−0.0007	−0.0008	−0.0008
	(0.0008	(0.0007	(0.0007)	(0.0005)	(0.0005)	(0.0005)
Matriculation in the following year						
All STEM tracks	−0.0017	−0.0016	−0.0015	−0.0012	−0.0012	−0.0012
	(0.0011	(0.0010	(0.0011)	(0.0010)	(0.0010)	(0.0010)
Selective STEM	-0.0025^{***}	-0.0023^{***}	-0.0022^{***}	−0.0002	−0.0002	−0.0002
	(0.0008	(0.0008	(0.0008)	(0.0006)	(0.0006)	(0.0006)
University STEM	0.0008	0.0006	0.0007	−0.0010	−0.0010	−0.0010
	(0.0010	(0.0009	(0.0009)	(0.0009)	(0.0008)	(0.0008)
University Medicine	0.0007	0.0007	0.0007	0.0002	0.0002	0.0002
	(0.0007	(0.0006	(0.0006)	(0.0010)	(0.0010)	(0.0010)
Vocational program	-0.0009^{**}	-0.0009^{**}	-0.0009^{**}	−0.0003	−0.0004	−0.0004
	(0.0004	(0.0004	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Selective Other	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005
	(0.0003	(0.0003	(0.0003)	(0.0004)	(0.0004)	(0.0004)
University Other	0.0005	0.0002	0.0002	−0.0003	−0.0004	−0.0004
	(0.0008	(0.0007	(0.0007)	(0.0009)	(0.0008)	(0.0008)
Not enrolled in Higher Ed	0.0008	0.0008	0.0008	0.0009	0.0009	0.0010
	(0.0008	(0.0009	(0.0009)	(0.0008)	(0.0009)	(0.0009)
Nb. observations	314, 389	344, 232	344, 227	282, 260	303, 206	303, 204

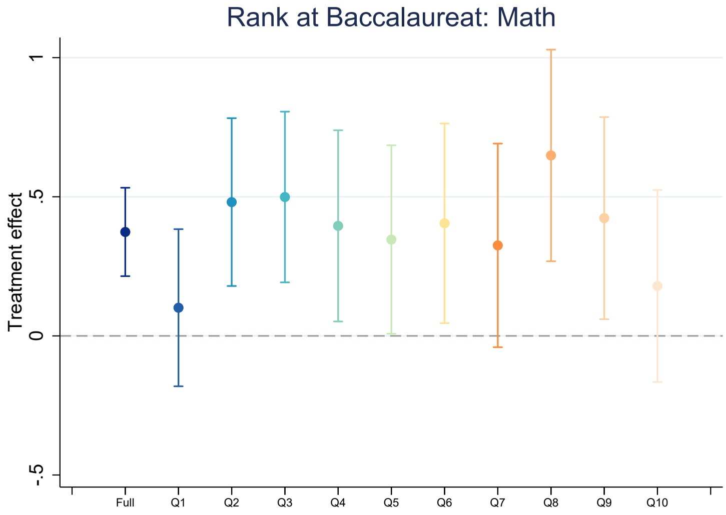
Each row reports the coefficients of the standardized *leave-one-out* teacher GDV obtained from the estimation of Equation 6 for the different outcomes listed on the first column. It is estimated on all the sample and separately for Grade 12 males and females. The regression includes high school, year and elective course 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 females in the classroom, and columns 3 and 6 control for the average leave-one-out GDV 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.

D.2 Additional Results: Heterogeneity by Initial Math Performance

Figure D1 – Impact of Teacher GDV on Math Performance at *baccalauréat* - By deciles of initial math performance



(a) Males

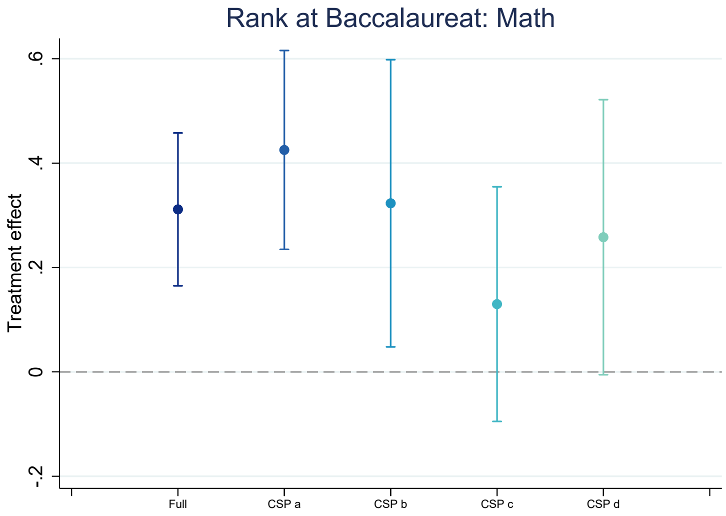


(b) Females

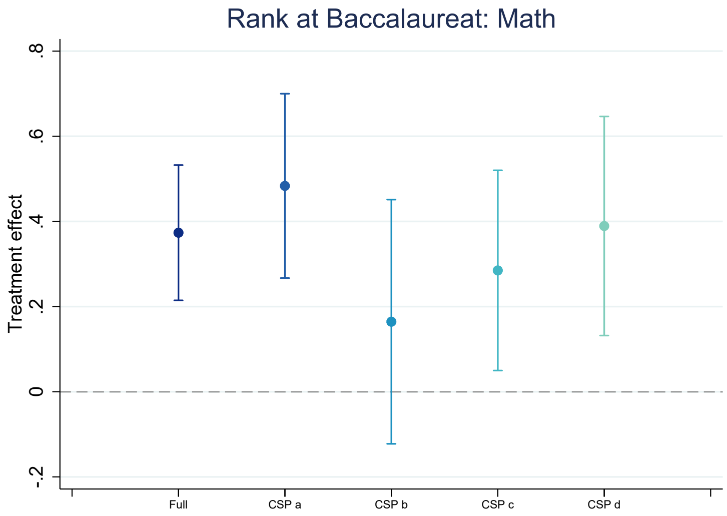
Notes: The figure reports the effect of a one standard deviation increase in teacher leave-one-out GDV on students' rank at *baccalauréat* in math separately by gender and by initial performance in math. Initial math performance is measured as deciles of percentile rank in math obtained at the DNB nation exam in Grade 9. The solid dots show the estimated coefficients, with 95 percent confidence intervals denoted by vertical capped bars. The coefficients are estimated on students for whom the high school × elective × year cell contains more than one math teacher.

D.3 Additional Results: Heterogeneity by Social Background

Figure D2 – Impact of Teacher GDV on Math Performance at *baccalauréat* - By Social Background



(a) Males



(b) Females

Notes: The figure reports the effect of a one standard deviation increase in teacher leave-one-out GDV on students' rank at *baccalauréat* in math separately by gender and by socioeconomic background. The solid dots show the estimated coefficients, with 95 percent confidence intervals denoted by vertical capped bars. The coefficients are estimated on students for whom the high school × elective × year cell contains more than one math teacher.

E Mechanisms: Estimation Details and Complementary Results

E.1 Estimating the Teacher Grading Bias

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

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

The grading bias assigned to class c is actually the average bias observed in any other classes taught by the same teacher except class c itself, i.e. it is the leave-one-out grading bias. A negative (positive) grading bias is indicative of a bias in favor of females (males).

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

Table E7 – Teachers' Average Math Grading Bias for Grade 12 Students

	Males			**Females**			**Teacher**
	G12 maths	Bac maths	Diff.	G12 maths	Bac maths	Diff.	**bias**
	(1)	(2)	(3)	(4)	(5)	(6)	
All teachers	−0.017	0.043	−0.060	0.020	−0.048	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	364,611	343,945		319,499	306,551		

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 males and females. 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 favour of girls.

E.2 Estimating the Teacher Value-Added

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

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

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

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

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

After the regression, we predict students' test scores residuals adjusted for observables.[A.4] Finally, for each teacher's class in year t, we compute the average test score residual. This should be seen as a proxy for teachers quality in the class taught in year t.

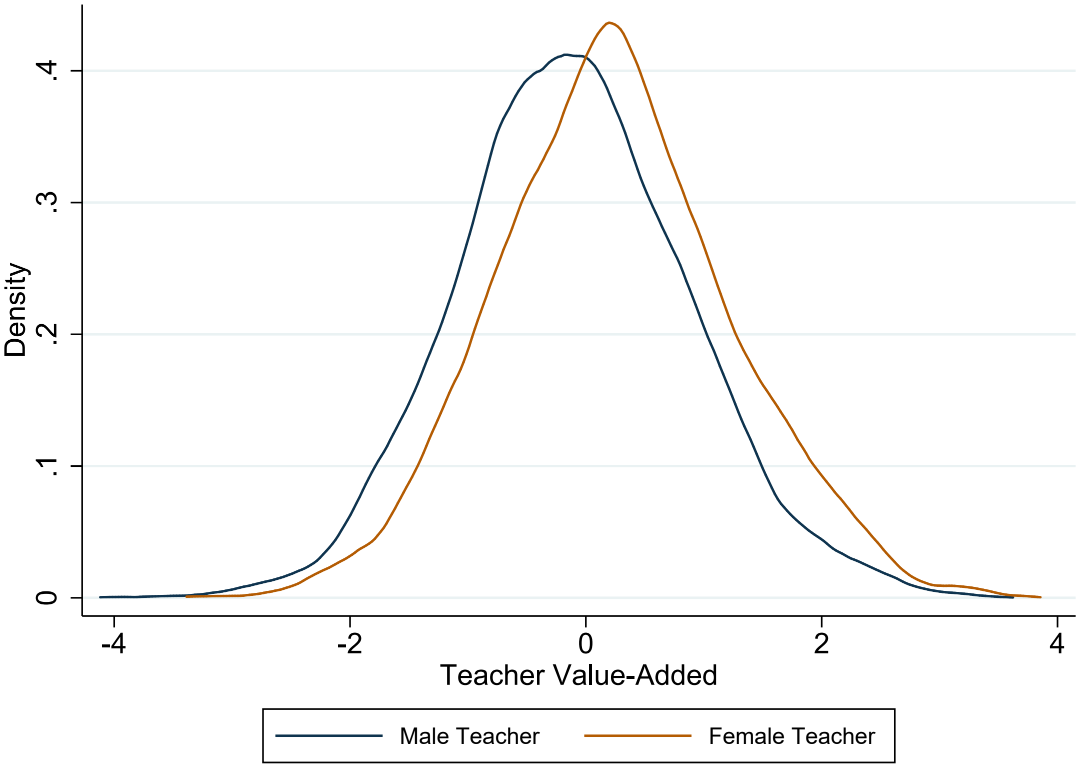
Step 2: Regressing teachers' quality in year t on its lags and leads. We regress the average test score residuals of teachers in t on those average residuals in years $t-1, t-2, \dots$ and $t+1, t+2, \dots$. The OLS coefficients obtained from this regression tell us how strongly current teacher performance 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 E3.

[A.4] 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 E3 – 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).

E.3 Complementary Results - Impact of Teacher GDV with Teacher Grading Bias or Value-Added controls

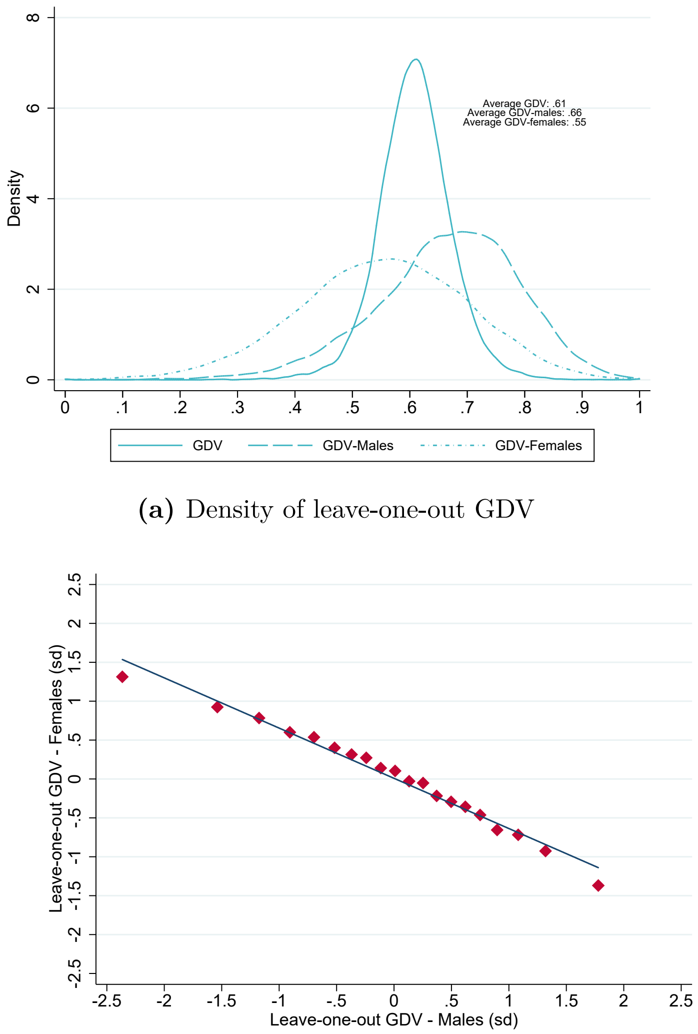
Table E8 – Impact of Teacher GDV - Mechanisms - All outcomes

	Males			**Females**		
	Grading bias	Value added	Value added	Grading bias	Value added	Value added
	(1)	(2)	(3)	(4)	(5)	(6)
Academic performance						
Rank at *baccalauréat*: math	0.3714***	0.3452***	0.2486***	0.3967***	0.4084***	0.2825***
	(0.0770	(0.0763	(0.0684)	(0.0851)	(0.0840)	(0.0727)
Rank at *baccalauréat*: philo	−0.0078	−0.0039	0.0033	−0.0011	0.0087	0.0135
	(0.0768	(0.0763	(0.0766)	(0.0809)	(0.0804)	(0.0807)
Type of STEM programs ranked first in the ROL						
All STEM tracks	−0.0029***	−0.0026**	−0.0027**	−0.0015	−0.0015	−0.0015
	(0.0011	(0.0011	(0.0011)	(0.0011)	(0.0010)	(0.0010)
Selective STEM	−0.0015	−0.0013	−0.0014	0.0003	0.0003	0.0003
	(0.0010	(0.0010	(0.0010)	(0.0008)	(0.0008)	(0.0008)
among which: biology	−0.0003	−0.0003	−0.0003	0.0000	0.0001	0.0001
	(0.0003	(0.0003	(0.0003)	(0.0004)	(0.0004)	(0.0004)
among which: math, physics	−0.0012	−0.0010	−0.0011	0.0002	0.0003	0.0002
	(0.0009	(0.0009	(0.0009)	(0.0007)	(0.0007)	(0.0007)
University STEM	−0.0015***	−0.0014**	−0.0014**	−0.0010	−0.0010	−0.0010
	(0.0006	(0.0006	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Vocational STEM	0.0003	0.0003	0.0003	−0.0010*	−0.0010*	−0.0010*
	(0.0008	(0.0008	(0.0008)	(0.0005)	(0.0005)	(0.0005)
Matriculation in the following year						
All STEM	−0.0022**	−0.0021*	−0.0023**	−0.0014	−0.0014	−0.0014
	(0.0011	(0.0011	(0.0011)	(0.0010)	(0.0010)	(0.0010)
Selective STEM	−0.0025***	−0.0024***	−0.0025***	−0.0004	−0.0004	−0.0005
	(0.0008	(0.0008	(0.0008)	(0.0007)	(0.0007)	(0.0007)
University STEM	0.0002	0.0003	0.0002	−0.0010	−0.0010	−0.0010
	(0.0010	(0.0009	(0.0010)	(0.0008)	(0.0008)	(0.0008)
University Medicine	0.0007	0.0007	0.0007	0.0002	0.0002	0.0002
	(0.0007	(0.0006	(0.0006)	(0.0009)	(0.0009)	(0.0009)
Vocational program	−0.0012***	−0.0011***	−0.0012***	−0.0003	−0.0002	−0.0002
	(0.0004	(0.0004	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Selective Other	0.0003	0.0003	0.0003	0.0005	0.0004	0.0004
	(0.0003	(0.0003	(0.0003)	(0.0004)	(0.0004)	(0.0004)
University Other	0.0004	0.0004	0.0004	−0.0001	−0.0003	−0.0004
	(0.0007	(0.0007	(0.0007)	(0.0008)	(0.0008)	(0.0008)
Not enrolled in Higher Ed	0.0020**	0.0018**	0.0020**	0.0008	0.0010	0.0012
	(0.0008	(0.0008	(0.0008)	(0.0009)	(0.0009)	(0.0009)
Nb. observations	340,729	342,380	342,380	300,020	301,409	301,409

Notes:

E.4 GDV by Gender: Distribution and Correlation

Figure E4 – Distribution and Correlation of Teachers Leave-one-out GDV by Gender



(a) Density of leave-one-out GDV

(b) Correlation

Panel (a) of this figure shows the distributions of math teachers' overall leave-one-out GDV, as well as the teacher accuracy computed for female students (*leave-one-out* GDV-females) and males students respectively (*leave-one-out* GDV-males). Panel (b) shows binned averages of GDV-males and GDV-females and plots the fitted regression line.