Teachers' Leniency: Experimental Evidence

Isaac Parkes * Alberto Prati † Etienne Dagorn‡

July 7, 2026

Abstract

Two teachers grading the same student can reach very different conclusions — yet little is known about which teacher characteristics drive such disagreement. We investigate this question through a large-scale online experiment in which over 1,600 French secondary-school teachers independently evaluated the same set of fictitious student transcripts, generating hundreds of ratings per transcript. This design allows us to decompose grade variation into a component attributable to the student (transcript fixed effects) and one attributable to the teacher (rater fixed effects). After controlling for transcript fixed-effects, we find that teacher leniency accounts for about 40% of the remaining grade variance. These differences are largely unrelated to teachers' gender, subject, teaching style, or job satisfaction, but are positively associated with altruism (measured via an incentivized dictator game) and exhibit a U-shaped relationship with age, with middle-aged teachers being the most strict. An incentivized task reveals that most teachers perceive themselves as more lenient than their colleagues - in particular women - but middle-aged teachers less so. Finally, teachers' stated tolerance for inter-rater disagreement comfortably exceeds the disagreement we actually observe — in contrast with findings from other professional domains. These results speak to the design of fair assessment panels and suggest that differential leniency, while substantial, may be partly predictable and therefore correctable.

Keywords: Teacher grading; Grading bias; Teacher leniency; Noise audit; Inter-rater disagreement; Experiment.

JEL codes: I21, C93, D91.

*London School of Economics, London, United Kingdom.

†University College London, London, United Kingdom; University of Oxford, Oxford, United Kingdom; London School of Economics, London, United Kingdom.
Corresponding author, email: a.prati@ucl.ac.uk

‡Univ. Lille, CNRS, UMR 9221 LEM Lille Économie Management, 42 rue Paul Duez, F-59000, Lille, France - etienne.dagorn@univ-lille.fr

1 Introduction

Grades play a pivotal role in shaping students' academic trajectories and, ultimately, their labor market prospects. They guide access to selective tracks, influence postsecondary opportunities, and serve as key signals of ability for students, families, and institutions. Yet, a large body of research in economics shows that grades and teacher evaluations are not neutral measures of student ability, and may be subject to biases.

Grading biases - intended as systematic deviations from a desirable benchmark - can be organized into three categories. First, *discriminatory biases*, i.e., differences driven by students' observable characteristics. Substantial evidence shows that, even among students of comparable ability, grades differ by gender (Lavy, 2008; Lavy and Sand, 2018; Protivínskỳ and Münich, 2018; Carlana, 2019; Terrier, 2020; Contreras, 2024), race and ethnicity (Gershenson et al., 2016; Quinn, 2020; Alesina et al., 2024; Bredtmann et al., 2024; Zhu, 2024), and socio-economic background (Doyle et al., 2023; Burgess and Greaves, 2013). Second, *ingroup biases*, i.e., favoritism or discrimination stemming from the matching between student and teacher characteristics. For instance, Feld et al. (2016) compare blind and non-blind grading and find significant ingroup effects based on nationality. Using a similar design, Lavy et al. (2022) shows that subtle cues of religious affiliation can trigger small ingroup favoritism.[1] The third and last source of grading bias is the one due to systematic differences in teachers' grading styles, i.e., their relative *leniency*. It relates to the "inner yardstick" used to map performance into grades. This source of bias has generally been under-studied in economics, and indeed little is known about how teacher's characteristics (such as their gender, age or tenure) predict the grades they give, holding students' characteristics constant.

In principle, differences in teachers' grading styles could be studied using observational data, but doing so is empirically challenging.[2] Most grades are assigned by a single teacher and often incorporate soft information (such as personal interactions or prior knowledge of students) that is traditionally unobserved. Even in settings where multiple raters assess the same candidates (such as admission committees) administrative data typically feature a small number of raters evaluating many items and contain limited or no information on raters themselves.

To overcome this limitation, we set up a large-scale vignette experiment with professional teachers. In collaboration with eight French regional school districts, we recruit more than 1,600 professional secondary-school teachers from both middle and high schools, who evaluate a set of ten fictitious transcripts, using a numerical grade (0-30). By holding student characteristics constant across raters, this design allows us to treat between-teacher variation in grading as fully exogenous and to purge our estimates of confounds arising from actual differences in

[1] Some assignment-based studies find benefits from being taught by a teacher of the same race or gender (e.g., Dee (2004); Carrell et al. (2010); Lusher et al. (2018); Gershenson et al. (2022); Dee (2007); Lim and Meer (2017); Winters et al. (2013); Gong et al. (2018); de Gendre et al. (2024)). However, these designs often cannot disentangle improved student performance from grading bias in assessments.

[2] In this paper, we use "grading style" and "leniency" interchangeably, to avoid repetitions. However, some other authors may use the concept of *grading style* to refer not only to leniency (the tendency to assign high grades) but also to moderateness (the tendency to avoid extreme grades).

student performance.[3] While our experimental setting abstracts from classroom interactions, our transcript evaluation task closely mirrors the information sets used in university admissions (part of the *Parcoursup* system, in France), where transcripts constitute a primary input.

Analysis of over 16,000 evaluations reveals that teachers account for about one fifth of the variance across grades, with grading style explaining about 40% of this effect. To probe teachers' awareness of this heterogeneity, we use an incentivized task in which respondents estimate both their own leniency relative to colleagues and ask them the level of disagreement they would consider acceptable between independent evaluations of the same transcript. Two findings emerge. First, teachers tend to perceive themselves as more lenient than their peers—a pattern consistent with motivated reasoning and self-serving attribution (Bénabou and Tirole, 2016). Second, the level of grading disagreement that teachers deem acceptable exceeds the variation we actually observe in our experiment. This second finding runs counter to the claim that evaluation systems are typically noisier than raters would consider acceptable (Kahneman et al., 2021), and suggest that teachers may have a more calibrated—and perhaps more tolerant—view of inter-rater variation than in other professions.

Grading styles are largely uncorrelated with teachers' gender, subject specialization, teaching practices, and job satisfaction. However, two characteristics are associated with leniency. First, lenient grades are associated with teachers' level of altruism - as measured in a real-incentive dictator game. For an extra €8 donated, teachers give on average a grade which is 0.1 standard deviations higher than their colleagues. Second, grading styles follow a pronounced U-shaped pattern with age: middle-aged teachers are systematically less lenient than both their younger and older colleagues. The implied distortions are economically meaningful: holding transcript characteristics constant, a student evaluated by a 40-year-old teacher receives a grade 0.08 to 0.21 SD lower than the same student evaluated by a 60-year-old teacher. This effect is comparable in magnitude to documented gender biases in math grading (Terrier, 2020), and three to seven times larger than the estimated grade penalty from exposure to a disruptive peer (Kristoffersen et al., 2015; Lavy and Yancu, 2025).

Why do we find these associations? On the one hand, one plausible interpretation of the positive association between teachers' generosity in the dictator game and their grading behavior is that both decisions reflect a stable, domain-general prosocial orientation. Grading can be viewed not only as a cognitive evaluation of performance, but also as a social allocation decision with material and symbolic consequences for students. More generous teachers may therefore exhibit stronger empathy, which, in turn, has been shown to be causally associated with altruism in economic interactions (Klimecki et al., 2016).[4] Our finding also connects with recent evidence

[3]Our design draws on the logic of correspondence and audit studies (Kline et al., 2022), adapted to the educational setting. Our identifying assumption is that teachers who are relatively lenient or strict with fictitious transcripts behave similarly with real ones—a response-consistency condition standard in the reporting-function literature (King et al., 2004; Kaiser, 2022).

[4]More broadly, prosocial dispositions have been shown to predict consequential decisions on the labour market (Kosse and Tincani, 2020). Workers reduce effort when higher productivity lowers coworkers' earnings (Bandiera et al., 2005), more altruistic individuals are more likely to sort into public-service jobs (Dur and Van Lent, 2018), and field evidence suggests that effort toward employers reflects warm glow and gift exchange more than pure

that teachers tend to assign higher grades to identifiable recipients (Sahlström and Silliman, 2024), similarly to what is observed in dictator games, where dictators are more generous when recipients are not anonymous (Charness and Gneezy, 2008). On the other hand, some recent developments in the anchoring vignettes literature might help answer why we find a U-shape relationship between age and grades. For instance, Benjamin et al. (2023) interview over 3,000 people to estimate predictable differences in their general use of subjective scales. Participants are asked to make assessments on a 0-to-100 scale of items without physical units, but where the "true" latent answer should be the same across respondents.[5] They find no difference in the average response across genders, but document a U-shape in age, where middle-aged respondents tend to use the scale more stringently than the rest of the population. Our results are consistent with theirs.

This work relates to two strands of the literature. The first one is the large literature about biases in teachers' evaluations (see Diris et al., 2026, for a comprehensive review). Grading biases have been shown to affect track placement, field of study, persistence in education, and later labor market outcomes (Clotfelter et al., 2010; Lavy and Megalokonomou, 2019; Carlana, 2019; Carlana et al., 2022). Existing studies, however, focus on how *students*' characteristics drive grading differences, implicitly abstracting from systematic differences in grading styles when students' characteristics are held fixed. The *teacher perspective* has been approached in several distinct strands of literature. One strand studies teachers' contributions to student outcomes using value-added models (Chetty et al., 2014; Backes et al., 2024), teacher quality and evaluation (Carrell and West, 2010; Taylor and Tyler, 2012). Separarelty, the grade-inflation literature examines variation in grading standards and the consequences of more generous grading practices(Denning et al., 2026; Kezim et al., 2005; Filetti et al., 2010; Griffith and Sovero, 2021). Finally, the psychometric literature has long documented rater leniency and inter-rater disagreement in the evaluation of written work (Myford and Wolfe, 2003; Eckes, 2017). This "grading lottery", where some students are randomly assigned to a relatively lenient or strict rater, has been shown to affect standardized tests (Dhawan et al., 2012; Huang, 2012; Huang and Whipple, 2023; Huang et al., 2023). We complement this evidence by studying *who* grades differently, and their own beliefs about it.

Our work also connects with the growing literature on the role of noise and rater-specific distortions in high-stakes settings (Kahneman et al., 2021). Previous studies have looked at how aggregation rules and committee design affect outcomes in hiring panels (Engel et al., 2010; Bagues and Perez-Villadoniga, 2012; Hoffman et al., 2018), admissions committees (Bagues et al., 2017; Deschamps, 2023), peer review (Boudreau et al., 2016; Pier et al., 2018), and judicial decisions (Danziger et al., 2011; Kleinberg et al., 2018). Institutional designs can help achieve fairer outcomes (Kates et al., 2023), but require knowledge of who the raters are and

altruism (DellaVigna et al., 2022).

[5]The calibration questions include both visual anchors - e.g. "how dark is this circle?" [0-100] - and descriptive anchors - e.g., "Your friends joke about how forgetful you are. You do often get lost, even in places you should know well. It's difficult for you to remember names, and you often lose track of things. But you usually remember to set reminders for important appointments. If this situation described your life during the past year, how would you rate your ability to remember things?" [0-100].

how they differ. We contribute to this agenda by estimating latent and observable predictors of grading styles in an educational context. A natural policy response to our documented U-shape relationship with age is to rely on age-diverse committees, that combine mid-career teachers with junior and/or senior colleagues.

Several limitations of our approach should be acknowledged. First, our design identifies grading styles only in relative terms: in the absence of an external benchmark for the "correct" evaluation, we treat the mean of many independent assessments as the relevant reference point and study deviations from that consensus. This feature is standard in the literature on rater effects—for instance among physicians (Currie and MacLeod, 2017) and judges (Dobbie et al., 2018)—and parallels some previous research in education, where teacher value-added is identified relative to the performance of students assigned to other teachers (Chetty et al., 2014). Second, evaluations of fictitious transcripts carry no real consequences and may therefore be noisier than in high-stakes contexts, potentially attenuating our estimates. Third, the cross-sectional nature of the data prevents us from separating age effects from cohort effects, and selective attrition may further bias observed age patterns, if teachers with certain grading styles are more likely to exit the profession (Wiswall, 2013). Finally, our analysis focuses on transcript-based evaluations relying on hard information; grading styles may operate differently for more subjective tasks such as essays or oral examinations. Despite these limitations, the experiment offers a rare opportunity to simulate a "grading lottery" where the same student is assigned to hundreds of different raters who are subsequently interviewed.

The remainder of the paper is organised as follows. Section 2 describes the design and procedures, while Section 3 outlines our empirical strategy. Section 4 reports our results, and Section 5 concludes.

2 Design and procedures

2.1 The experiment

The study was presented to participants as an investigation of teaching practices and took approximately 20 minutes to complete. The study consisted of five sequential blocks, covering background information (Block 1), the core grading task (Block 2), implicit association measures (Block 3), incentivized redistribution choices (Block 4), and perception and belief surveys (Block 5). Blocks 2, 4 and 5 are the primary focus of this paper; Block 3 and the ingroup/gender components of Block 4 are analyzed in companion work (Monnet et al., 2025).

Block 2: Evaluation task. Teachers are instructed to take the role of a homeroom teacher and evaluate each transcript as they would during a class council meeting—a formal committee in French secondary schools where teachers collectively review students' progress and decide on promotion and support. This block always appeared first in the survey to rule out any contamination from later stages of the experiment. Each teacher evaluates 10 out of 16 fictitious

10th-grade student transcripts, randomly drawn and presented in random order to prevent systematic order effects.

We provide teachers with the key information typically available in this evaluation setting: student's name, grades in five core subjects (French, Math, Life and Earth Sciences, English, History), and indicators of school behavior (number of absences and late show-ups). We include school behaviors as teachers are required to account for this dimension in their evaluations.[6] To avoid unintended signals about students' socio-economic or ethnic background, names were selected through a preliminary survey in which trainee teachers rated common French first and last names by perceived socio-economic background. We then paired first and last names receiving similar ratings, ensuring that names carry no differential signal about social origin. The 16 transcripts correspond to all possible combinations of four binary dimensions: gender (male/female), behavior in class (poor/high), performance in STEM (poor/high), and humanities (poor/high). Table ST-2, in the appendix, summarizes the characteristics of each of the 16 transcripts. Figure 1 displays the screenshot of the user interface showing an example of the fictitious student transcript.

Teachers make three assessments that closely mirror real-world grading practices in French secondary schools. Specifically, each teacher makes three decisions:

1. Assign a numerical grade, reflecting overall school performance and school behaviour , which is a commonly studied outcome in the literature about grading biases (e.g., Lavy, 2008; Lavy and Sand, 2018; Terrier, 2020)[7];

2. Assign a verbal mark (*mention*) ranging from the lowest (worrisome record) to the highest (excellence). This outcome is of independent interest: because teachers must map a continuous internal judgment onto a small number of discrete categories, small differences in how they interpret category boundaries can translate into systematically larger grading disparities than those observed on a continuous scale. For completeness and robustness, we replicate all main analyses using this outcome in Appendix 3;

3. Choose a comment on the student's potential for progression among 4 options. We do not analyze this outcome, since response options were varied from transcript to transcript and do not have a clear ordinal structure (these responses are analyzed in a companion paper, Monnet et al., 2025).

Block 4: Dictator game In Block 4, participants took part in a standard dictator game. They were asked to allocate an endowment of 50 tokens (i.e., €50) between themselves and an

[6]The national guidelines regarding the content of school transcripts (called "*livret scolaire*") can be found at the following link. Annexe 3, article 8 refers to how school life aspects should be accounted for. https://www.education.gouv.fr/bo/16/Hebdo3/MENE1531425A.htm

[7]Teachers report their overall assessment on a 0–30 scale—rather than the standard French 0–20 scale used for transcript grades—to discourage them from simply averaging the transcript grades. A minority of teachers (n = 42) nonetheless appeared to interpret the scale as 0–20; we rescaled their responses by a factor of 1.5, and show in Appendix Table ST-5 that excluding these observations yields similar results.

anonymous random participant.[8] They were informed that 100 participants will be chosen at random and their decisions will be implemented. The share allocated to the other participant serves as our measure of altruism and is included to examine whether general prosocial preferences predict leniency. After the standard dictator game, participants also played four dictator games with identifying information about the recipient, to elicit ingroup and gender biases, but these tasks are not analyzed here (see instead Monnet et al., 2025).

Block 5: Perceptions and Beliefs The last block of the experiment consists in a survey to collect additional information on teachers' perceptions, attitudes, and beliefs (see Appendix Table ST-1 for details). These included:

(i) *perceived leniency*, i.e., teachers' perception of how lenient their grading is relative to that of their colleagues. Teachers answered the following question: *"At the beginning of the questionnaire, you assigned scores from 0 to 30 to different transcripts. In your opinion, what would be the average difference between the score you gave to one of these transcripts and the score one of your colleagues would have given?*". Belief elicitation was incentivized. One transcript was selected at random, and if the teacher correctly guessed the difference between their score and the average score given by their colleagues (within ±0.25 points), they would enter a draw to receive an additional €10. This measure allows us to examine whether teachers are aware of their own grading tendencies and whether self-perception aligns with actual behavior.

(ii) *acceptable deviation*, i.e., the answer to the question *"In your opinion, what difference between the two grades would you consider acceptable to ensure a fair and equitable process?"*. This allows us to assess whether teachers' normative standards for grading consistency are in line with the actual dispersion observed in our experiment.

(iii) their *job satisfaction*, measured using a standard 0–10 subjective well-being scale (inspired by De Neve and Ward, 2025): "*To what extent are you satisfied with your job, in general?*". This measure is included to test whether teachers who are more satisfied with their job invest more care in grading, and thus exhibit less leniency or more consistency.

(iv) their relative *self-efficacy*, i.e. their perceived ability to positively influence student progress compared with colleagues (inspired by Bandura and Adams, 1977). Self-efficacy was measured with a 1–7 agreement scale based on the statement:"*I am better than most of my colleagues in making students progress*". This measure is included to examine whether teachers who perceive themselves as more effective than their peers also grade more strictly, reflecting higher academic standards or greater confidence in their judgment.

[8] Each participant was matched with a volunteer from AFEV (*Association de la Fondation Étudiante pour la Ville*), an NGO providing homework support to students in disadvantaged neighborhoods in France (https://afev.org/). This matching was chosen to reflect a naturalistic setting in which teachers allocate resources to a student-facing beneficiary.

(v) their modern vs. traditional teaching style, based on an index adapted from Alan et al. (2018). See Appendix 4 for details about the questions. This measure is included to test whether teachers with more progressive, student-centered approaches tend to grade more leniently, consistent with a broader emphasis on encouragement over evaluation.

(vi) their mindset about the role of effort vs talent in academic achievement, based on an index adapted from Copur-Gencturk et al. (2021). See Appendix 4 for details about the questions. This measure is included to examine whether teachers who attribute academic success primarily to effort (rather than innate ability) grade more generously, rewarding perceived motivation over demonstrated performance.

2.2 Implementing the experiment and sample

Data Collection We contacted eleven school districts (*académies*), and eight agreed to take part in the study.[9] In October 2023, the school district representatives sent an email to the relevant teacher mailing lists in their districts, or to those under their subject-specific remit. The email included a hyperlink to our survey experiment. Although teachers were invited to take part in the survey on a purely voluntary basis, they were informed of the potential gains at the very beginning of the survey.

Inclusion criteria The experiment was programmed using oTree (Chen et al., 2016), and data were collected from November 2023 to January 2024. The median time for completing the transcript task (block 2) was 8 minutes. Encouragingly, the time allocated to evaluating each manuscript remains relatively stable from the first transcript to the last (see Appendix, Figure SF-3), thus suggesting sustained effort. To reduce noise, and in accordance with the pre-registration, we exclude (i) straight-liners and (ii) those who spent too little time on the transcripts. Straight-liners are defined as those who give the same numerical grade - rounded to the closest integer - to all transcripts. The response time threshold was defined as the bottom 1% of the response time distribution, i.e. 17 seconds on average per transcript. Results turn out to be largely unaffected by the removal of these outliers. Over 5,000 teachers clicked on the survey, and 1,922 completed it. Among them, 59 did not pass at least one of the two attention checks and were dropped. In addition, 10 straight-liners were identified and dropped, and 21 participants were removed because of their short response time. Out of these 1,832 participants, 10% were assigned to a treatment for a companion paper,[10] while 90% were assigned to our main experimental design, where they evaluated ten transcripts with the characteristics outlined in Appendix Table ST-2. These represent our analytical sample, which is made up of 1,644 participants, who reported as many measures of perceived leniency, and assigned 16,440 numerical grades.

[9] Participating districts: Amiens, Bordeaux, Dijon, Lille, Nantes, Nice, Rennes, and Toulouse. One district declined due to concerns about monetary incentives.

[10] The treatment was a non-incentivized version of the experiment that focused on gender bias. Eight transcripts were presented in a blind format, i.e., without any name or gender information. Results are published in Monnet et al. (2025).

Table 1: Teachers' Summary Statistics - Analytical Sample vs. National Population

	Participants		Population	
Variable	Mean	S.d.	Mean	Δ Mean
Age (years)	44.9	8.9	44.9	0.00
Female (%)	0.63	0.48	0.59	0.04
Experience (years)	18.2	9.2	16.8	1.4
STEM teacher (%)	0.41	0.49		
Middle school (%)	0.49	0.50	0.50	-0.01
Academic High school (%)	0.43	0.50	0.26	0.17
Middle and high school (%)	0.08	0.27		
Observations	1,644			

Notes: *Notes:* This table compares the observable characteristics of teachers in the analytical sample with those of the national population of public secondary-school teachers in France. National population statistics are taken from MENJS (2021), which reports means but not the associated standard deviations. The reported population shares do not sum to 100% because vocational-school teachers are excluded from these categories; they represent approximately 24% of the overall teacher workforce.

Sample characteristics and representativeness Given that participation was on a voluntary basis, we check for sample representativity by comparing our participants' demographics with those of the national teaching workforce using the "*Base Statistique des Agents*". Table 1 shows that participants are on average 44.9 years old, perfectly aligning with the population proportion. 63% of participants are female, slightly above the population average of 59%. The main difference between the two samples is the proportion of high school teachers compared to middle school teachers, which is about 1:1 in our sample compared to just 1:2 at the national level. Additionally, participants have slightly more experience (18.2 years vs. 16.8 years). Overall, the table highlights that the study sample closely resembles the general teacher population in key demographics.[11]

3 Empirical strategy

3.1 Definitions

An intuitive way to think about leniency is that it functions like a personality trait (Kahneman et al., 2021). Teachers can be arranged on a scale ranging from very strict to very lenient, just as a personality tests can measure the degree of neuroticism or extraversion. Like other traits, leniency is likely to be correlated with (i) demographic characteristics, (ii) relevant experiences and (iii) other behavioral traits - i.e., the focus of this paper.[12] Leniency may be conscious — for

[11]Of course, we can only check representativity on observable characteristics but cannot assess whether our sample differs on dimensions like a stronger commitment to advancing educational knowledge, which may also correlate with less gendered assessments.

[12]About differential leniency in the judicial system, Kahneman et al. (2021) explains that one can "arrange judges on a scale that ranges from very harsh to very lenient, just as a personality test might measure their degree of extraversion or agreeableness. Like other traits, we would expect severity of sentencing to be correlated with genetic

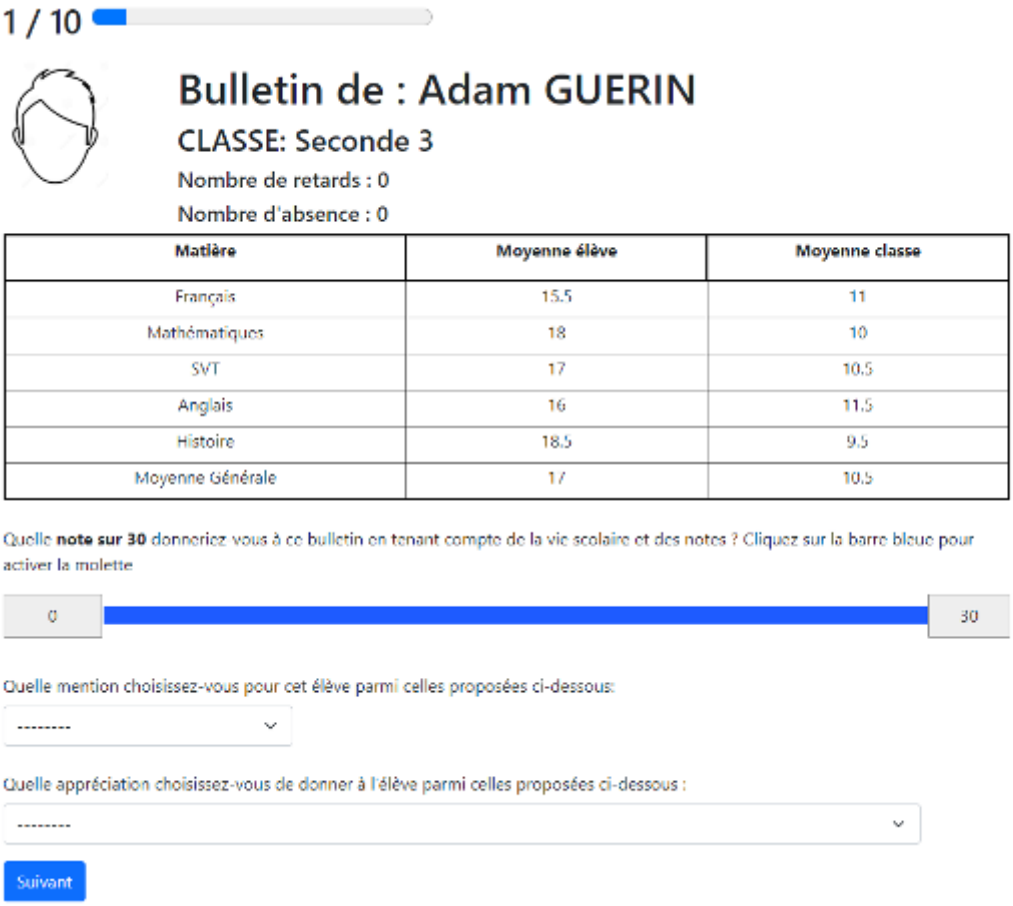


Figure 1: Screenshot of a transcript

instance, when a teacher deliberately inflates grades to avoid conflict or to appear popular — or unconscious, reflecting internalized standards that the teacher has never examined and might revise if confronted with evidence of how they compare to peers. This distinction matters for the self-reported measure of leniency we introduce below, which captures only the conscious component. Note that leniency, as defined here, is unconditional: it reflects a teacher's general tendency to grade high or low, irrespective of the characteristics of the student being evaluated. This distinguishes leniency from other forms of grading bias, such as differential treatment based on student gender or ethnicity.

Formally, by *leniency* (or *grading style*) we mean the discrepancy between the evaluation given by a teacher to an item and the average evaluation of that same item by other teachers. In our estimations, we implicitly consider the mean of the hundreds of independent evaluations as the "correct" evaluation, since there is no objective way to determine what the correct evaluation of a transcript should be. This principle is based on the well-documented wisdom of the crowds' effect, and it has been previously employed for assessing leniency, including in institutional settings (U.S. Sentencing Commission, 1991).

A useful way of modelling leniency is via a latent variable model. Let y_{it} be the grade

factors, with life experiences, and with other aspects of personality. None of these has anything to do with the case or the defendant."

reported by teacher t to item i, and y_i^* the latent true score of item i, i.e., the average evaluation of the item (here, a transcript) by a large number of independent raters. We can model the relationship between the reported grade y_{it} and the true score y_i^* as:

$$y_{it} = y_i^* + \underbrace{l_t}_{\text{leniency}} + \underbrace{n_{it}}_{\text{noise}} \tag{1}$$

Where l and n follow some distributions both centered at zero. Here, parameter l_t captures teacher's *leniency*, i.e., the deviation of teacher t's evaluation from their average colleague. Parameter n_{it} captures *noise*, i.e., a teacher being particularly lenient or strict on a specific transcript, due to idiosyncratic biases, low effort or randomness. In the literature on measurement error, the systematic error captured by parameter l_t is akin to the level of *validity* of teacher t's grading style, while the random component n_{it} is akin to its *reliability*.[13] We further assume that leniency l_t and noise n_{it} are independent—that is, a teacher's systematic grading bias is unrelated to their item-specific deviations—so that, for example, lenient teachers are not also more variable in their evaluations across transcripts. We assume that l_t is constant across items — that is, a lenient teacher is equally lenient regardless of which transcript they evaluate. This assumption would be violated if, for example, a teacher's leniency depended on the perceived quality of the work being evaluated, being more lenient toward weaker students.

By definition, the average evaluation is accurate, as reflected by the fact that l_t and n_{it} are both centered at zero. While both components have zero mean by construction, their variances are empirical quantities of interest: the variance of l_t captures how much grading standards differ across teachers, while the variance of n_{it} captures how inconsistently a given teacher applies their own standards across items.[14] However, for each singular item, there is a certain amount of undesirable variability in the evaluations. Since, usually, only one teacher evaluates the same item, this variability goes undetected.

Importantly, the unfairness introduced due to some teachers being more or less lenient does not cancel out. For example, if two equivalent transcripts both deserve 25/30, but one is evaluated by a lenient teacher who gives 27/30 and the other is evaluated by a strict teacher who gives 23/30, the average is still 25/30, but the grading system would be unfair because one student gets more than they deserve and the other gets less. This constitutes a form of inequity: students with identical ability receive different grades purely as a function of which teacher happens to evaluate them, a source of unfairness that is invisible when only aggregate statistics are examined. In Appendix 2, we offer an example to help fix ideas for readers who are unfamiliar with this type of decomposition.

[13] Our approach is anchored in Item Response Theory. In the vocabulary of Kahneman et al. (2021, chap.6), parameter l_t corresponds to "level noise", while n_{it} corresponds to "pattern noise".

[14] Leniency is estimated from only 10 transcripts per teacher, which introduces sampling error. Intuitively, a teacher's estimated leniency $\hat{l}_t$ is their average score minus the mean. Even a teacher with no systematic bias will appear slightly lenient or strict by chance, depending on which particular transcripts they happened to evaluate. This means that the variance of $\hat{l}_t$ across teachers reflects both genuine differences in leniency and pure sampling noise, and the two cannot be perfectly separated with only 10 observations. As a result, our decomposition may overstate the contribution of leniency relative to idiosyncratic noise.

3.2 Identification

To estimate which characteristics of the respondents predict leniency we use two different estimation strategies: (i) fixed-effects ordered probit; and (ii) fixed-effects OLS. For each regression, we will correct for multiple-hypotheses testing and assess statistical significance using False Discovery Rate correction (q < 0.05).

3.2.1 Fixed-effects ordered probit

We can model subjective evaluations as the outcome of a two-step process. In the first step, the teacher observes the characteristics of item i and forms an internal evaluation of the student (the latent variable). In a second step, the internal evaluation is mapped into the 0-30 grading scale (the observed score). We call this mapping the "reporting function" (see Oswald, 2008), and use a semi-parametric model similar to Kaiser (2022) and Seré and Decancq (2025) to estimate leniency in the reporting function.

Let's denote $\widetilde{y_{it}}$ the internal (latent) score formed by teacher t about item i (a transcript, in our experiment), i.e., the outcome of the first step of the cognitive process. Let's denote y_{it} the (observable) score expressed for item i by teacher t, i.e., the outcome of the second step of the cognitive model.

Assume that the first step of the model can be specified as follows:

$$\widetilde{y_{it}} = Z_t\alpha + \beta_i + \gamma e_{it} + \varepsilon_{it} \tag{2}$$

where the matrix Z_t includes the teacher's characteristics. The vector β_i corresponds to transcript fixed effects that capture transcript-specific quality. These fixed effects account for the fact that not all teachers evaluate the same transcript. The term e_{it} is the transcript-specific response time, which is intended to proxy effort. The error term ε_{it} captures idiosyncratic deviations reflecting taste differences (e.g., weighting STEM results more than humanities) or measurement error. For identification of the ordered probit model, we assume $\varepsilon_{it} \sim \mathcal{N}(0, 1)$. In the estimation, standard errors are clustered at the teacher level to account for the within-teacher correlation in the noise component ε_{it}.

In the second step of the cognitive model, the individual's internal evaluation score $\widetilde{y_{it}}$ is translated onto the observed grading scale through a reporting function. This mapping captures how respondents convert their internal, latent evaluation into a discrete observed score. Specifically, the internal score is compared to a set of K ordered thresholds corresponding to the levels of the grading scale. For each response category $k = 1, \ldots, K$, the observed score y_t is determined according to:

$$y_t = k; \Leftrightarrow; \tau_t^{k-1} < \widetilde{y_{it}} \leq \tau_t^k, \tag{3}$$

where the thresholds satisfy $-\infty = \tau_t^0 < \tau_t^1 < \cdots < \tau_t^{K-1} < \tau_t^K = +\infty$, ensuring that all possible values of the internal evaluation are mapped to a valid category on the (bounded) grading

scale.

Each reporting function is fully described by a vector of $K-1$ finite thresholds, $\tau_t^1, \ldots, \tau_t^{K-1}$. The absolute values of these thresholds are not intrinsically meaningful; rather, their interpretation depends on the scale of the internal evaluation $\widetilde{y_{it}}$. That is, they define the boundaries between observed categories relative to the latent evaluation rather than representing standalone quantities.

We assume that threshold values vary only with the teachers' characteristics. That is, for each $k = 1, \ldots, K-1$, we assume that:

$$\tau_t^k = \tau^k + \delta Z_t \tag{4}$$

where the matrix Z_t identifies the teacher's characteristics. Thus, all grades from a teacher with the combination of J characteristics $Z_t \equiv (z_{1t}, \ldots z_{Jt})$ are assumed to share a common reporting function. The contribution of each observation to the log-likelihood of the ordered probit model is then given by:

$$\mathcal{L}_{it} = \sum_{k=1}^{K} \mathbb{1}(y_{it=k}) ln[\Phi(\tau^k + \delta Z_t - \beta_i - \alpha Z_t - \gamma e_{it}) - \Phi(\tau^{k-1} + \delta Z_t - \beta_i - \alpha Z_t - \gamma e_{it})] \tag{5}$$

Here, $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, and $\mathbb{1}(\cdot)$ denotes the indicator function. Throughout, we maximize:

$$\mathcal{L} = \sum_{i=1}^{N_t} \sum_{t=1}^{T} \mathcal{L}_{it} \tag{6}$$

to estimate the parameters. N_t denotes the number of observations for teacher t (10, in our sample), and T is the total number of teachers (1,644 in our sample). The thresholds (and hence the mapping from internal evaluation to response) can shift across teachers, thus identifying leniency.

In practice, though, the estimator does not converge. A problem we failed to foresee is that the highest-quality transcripts (n. 1, 2, 9, and 10) receive grades bunched at the top of the scale, preventing the ordered probit estimator from converging and estimating the thresholds (see the distribution of grades in Figure 2A, as well as SF-1 and SF-2). We therefore omit data from transcripts 1, 2, 9 and 10 from those regressions.[15] To allow an estimation on the whole sample, we can move on to fixed-effects OLS.

3.2.2 Fixed-effects OLS

If we treat the grades as cardinal values, we can estimate the model specified in equation (2) directly by linear least squares, by imposing $\widetilde{y_{it}} \equiv y_{it}$, that is:

$$y_{it} = Z_t \alpha + \beta_i + \gamma e_{it} + \varepsilon_{it} \tag{7}$$

[15] Although numerical grades could be given in 0.25 increments, we round the grades to the closest integer scale since we do not have enough statistical power to estimate 120 cutoff points ($\{ x \in [0, 30] \mid x = 0.25 \cdot k,\ k \in \mathbb{Z},\ 0 \leq k \leq 120 \}$).

As in the case of ordered probit, we identify α_z from the variation in Z_t across teachers who grade the same transcript. If a component z_t of the matrix Z_t - say, "tenure" - is associated with a positive coefficient α_z, it means more experienced teachers systematically give higher grades than the transcript average — i.e., tenure predicts leniency. The term e_{it} is the transcript-specific response time, which is intended to proxy effort. As for ordered probit, we cluster the standard errors at the teacher's level and assume ϵ_{it} to be normally distributed and centered at zero.

4 Results

4.1 Grading styles and their contribution to grading variance

Figure 2 shows within-transcript and between-teacher variation in grading, meaning that not all teachers assign the same grade to a given transcript, and that grading styles are not homogeneously distributed across teachers. Panel A reveals spread in grades within each transcript: on average, the same transcript receives grades with a standard deviation of 2.62 points and a mean interquartile range of 2.77 points (out of 30). Panel B confirms this pattern at the teacher level: the distribution of estimated teacher fixed effects spans nearly 14.6 points (from 21.0 to 35.6), with a standard deviation of 1.57 points, indicating that some teachers are systematically more lenient or more strict than average.

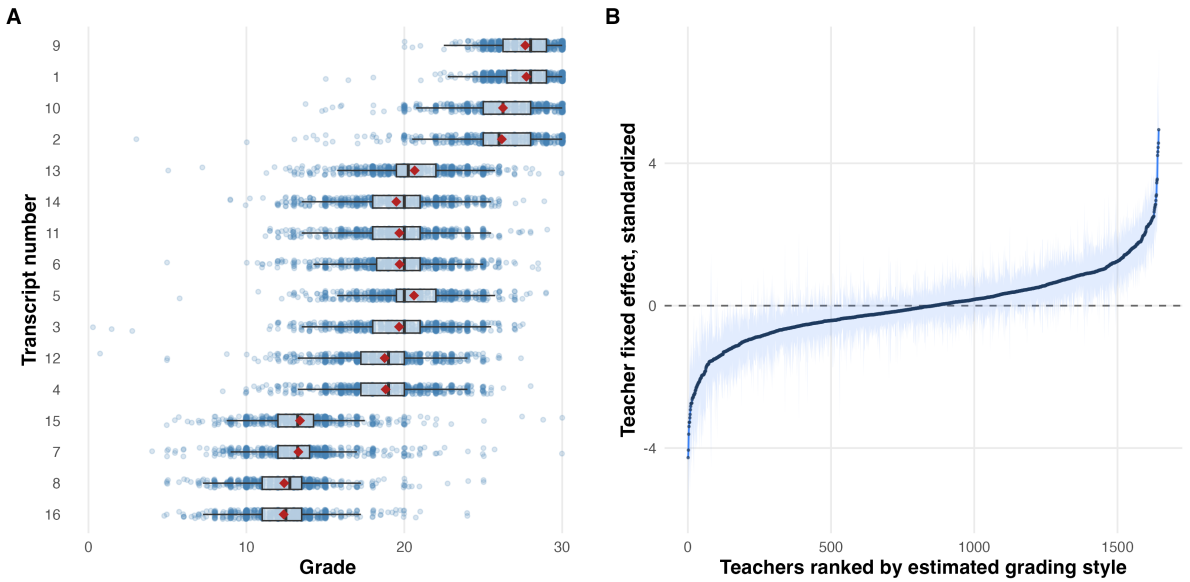


Figure 2: Grade variation across transcripts and teachers

Notes: Panel A shows the distribution of grades across transcripts, ordered by median grade. Each point represents an individual grade ; box plots display the median and interquartile range; red diamonds indicate the mean. Panel B shows teacher fixed effects estimated from a regression of grades on transcript and teacher indicators, standardized to have mean zero and unit standard deviation. Teachers are ranked from most lenient to most strict. The shaded band reports approximate 95% confidence intervals based on within-teacher residual variation after controlling for transcript fixed effects. Both panels draw on the same dataset of 16,440 grades assigned by 1644 teachers across 16 transcripts.

Figure 3 displays the distribution of Mean Squared Error (MSE) — the squared difference between the grade assigned by a teacher and the consensus grade derived from the crowd — decomposed into leniency and noise components (see eq.1). On average, leniency accounts for about 40% of total MSE, while noise represents the remaining 60%. However, this aggregate decomposition masks important heterogeneity across the MSE distribution. Among teachers in the bottom decile, leniency accounts for just 16.9% of the MSE, while it accounts for 51.5% among those in the top decile.

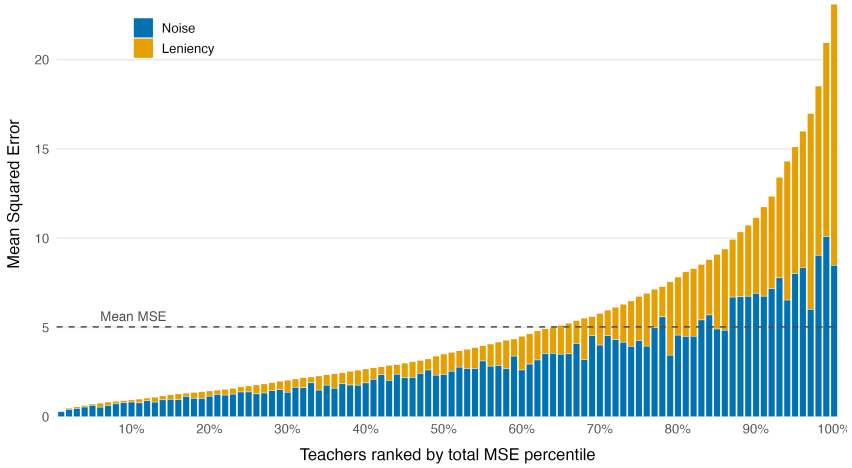


Figure 3: Mean MSE by Leniency Share Percentile

Notes: Bars represent deciles of teachers ranked by their leniency share of total MSE (prediction error). Leniency captures systematic grading bias; Noise captures idiosyncratic variation. The dashed line indicates the overall sample mean MSE.

Table 2 offers a formal statistical test of the importance of teachers' effects, based on an ANOVA decomposition. The total grade variance into three components. *Transcript effects* account for four-fifths (80.1%) of variance in grades, indicating that teachers largely respond homogeneously to differences in transcript quality when assigning grades. *Teacher effects*, representing systematic differences in grading standards (leniency), account for 8.2% of variance. The *residual* 11.7% is attributed to noise and non-systematic differences. The F-test shows that between-teacher variations are highly significant determinants of the assigned grades.

Overall, teachers account for about one fifth of the variance in evaluations, with grading style explaining about 40% of this effect (i.e. 8.2 percentage points out of 19.9). This is what we refer to as the "grading lottery": a substantial portion of the grade is not explained by the quality of the transcript, hence generating unfair differences across students who would deserve the same grade but have the (un)luck of being assigned to a specific teacher.

Result 1: Teachers account for about one fifth of the variance in evaluations, with grading style explaining about 40% of this effect.

Table 2: Detailed Analysis of Variance Results: Effects of transcript and teacher on numerical grade.

Source of variance	Unique variance	Sum of squares	Mean squares	F-test ratio	Significance
Transcript	80.1%	398,723	26,582	6730	<0.001
Teacher	8.2%	40,983	25	6.3	<0.001
Residual	11.7%	58,381	4		

Notes: This table decomposes the total variance in transcript grades into three components. The 'Transcript' component (80.1%) represents systematic differences in quality between transcripts. The remainder (19.9%) is decomposed in two components. The 'Teacher' component (8.2%) captures systematic differences in grading standards between teachers (level noise). The 'Residual' component (11.7%) represents the combined effect of teacher-transcript interactions and random noise. Analysis based on 16,440 grades from French teachers evaluating 10 student transcripts. The F-test shows that both transcript and teacher effects are highly significant determinants of numerical grades.

4.2 Teacher's *beliefs* about their grading style and acceptable grading variance

We now present descriptive evidence on teachers' beliefs about leniency and what they consider acceptable disagreement in their grading task. Figure 4 summarizes both dimensions and the gap between beliefs and actual benchmarks.

Panel (a) plots the distribution of teachers' perceived leniency. On average, teachers believe themselves to be about 2 scale points (on a 0-30 scale) more lenient than their colleagues. About 3-in-10 believe themselves to be relatively strict, while roughly twice as many believe themselves to be relatively lenient. A t-test strongly rejects the null hypothesis that mean perceived leniency equals zero ($p<0.001$). Insofar as leniency is perceived as a form of kindness, this pattern is consistent with a well-documented tendency to rate oneself favorably on positive characteristics (Zell et al., 2020), and which is often explained as an instance of motivated reasoning (Bénabou, 2015; Epley and Gilovich, 2016).

Panel (b) turns to the maximum deviation teachers consider acceptable. The average response was 4.87 on a 0-30 scale. This is well *above* the average absolute deviation actually recorded in our experiment, which is 2.60 (st.dev. 0.38), and a t-test strongly rejects the equality of the two ($p<0.001$). That is, the recorded level of grade deviation is within what's considered acceptable by teachers themselves. This result stands in contrast with Kahneman et al. (2021)'s warning that the amount of undesirable variation in a decision process is often beyond what professionals would consider an acceptable threshold. In our setting, however, the opposite holds.

Result 2a: Teachers tend to think they are more lenient than their average colleague.
Result 2b: The recorded level of grade deviation is within what's considered acceptable by teachers themselves.

In the following two sections, we'll regress observed and perceived leniency on a set of demographic, professional and behavioral characteristics. For completeness, in the appendix

(table ST-6), we run a similar exercise for the Mean Absolute Deviation and its subjective counterpart, i.e., the maximum acceptable deviation.

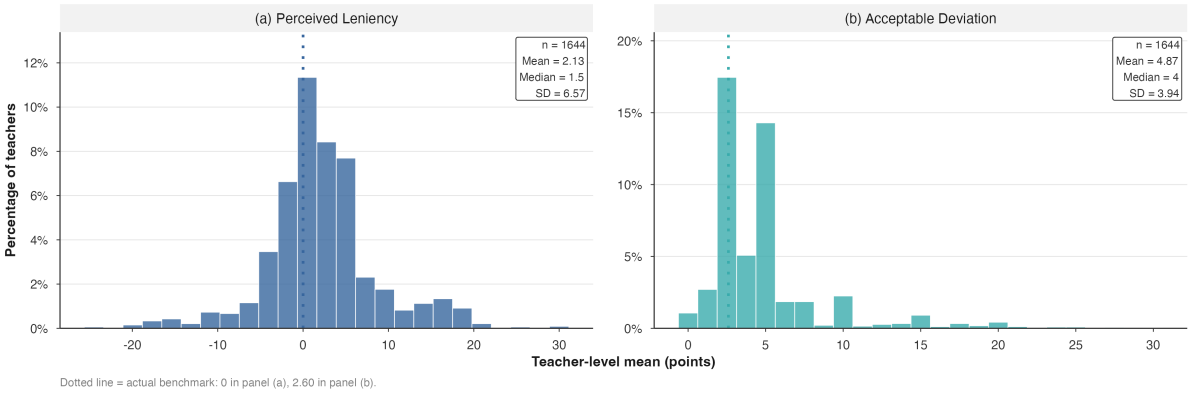


Figure 4: Teacher's beliefs about their grading style and acceptable grading variance

Notes: Panel (a) shows the distribution of teachers' perceived leniency relative to their colleagues. The dotted vertical line marks the actual mean leniency, i.e. zero (by construction). Perceived leniency is measured from the response to the question "In your opinion, what would be the average difference between the score you gave to one of these transcripts and the score one of your colleagues would have given?". Panel (b) shows the distribution of teachers' acceptable deviation from colleagues' grades. The dotted vertical line marks the actual average absolute pairwise difference (2.60). Acceptable deviation is measured from the answer to the question "In your opinion, what difference between the two grades would you consider acceptable to ensure a fair and equitable process?".

4.3 The predictors of teachers' leniency

In this section, we move on to study the empirical predictors of grading style and teachers' beliefs. We consider three blocks of characteristics: *i)* demographics (gender, age), *ii*) professional characteristics (tenure, subject), *iii*) behavioral characteristics (altruism, teaching style, job satisfaction, self-efficacy). Table ST-3, in the appendix, provides descriptive statistics for each predictor.

We use regression analysis and examine, in Table 3, any systematic variations in grading leniency which can be associated with the characteristics of the respondents. Columns (1)-(4) are estimated using a fixed-effect ordered probit, and use a subsample of transcripts ($N = 12,368$), due to the convergence problems explained in Section 3.2.1. Columns (5)-(8) are estimated using a fixed-effect OLS on the whole sample of transcripts ($N = 16,440$). The table shows the standardized coefficients estimated from the full model (columns (4) and (8)), and with each block of explanatory variables introduced incrementally (columns (1)-(3) and (5)-(7)). All models are transcript-level regressions where the dependent variable is the 0-30 numerical grade. The inclusion of transcript fixed effects implies that coefficients can be directly interpreted as conditional correlations with leniency. We correct for multiple hypothesis testing and use the Benjamini–Hochberg procedure to control for False Discovery Rate, with significance threshold q<0.05. Surviving significant coefficients after the correction are reported in bold.

Results show that leniency is primarily associated with two characteristics: teachers' age

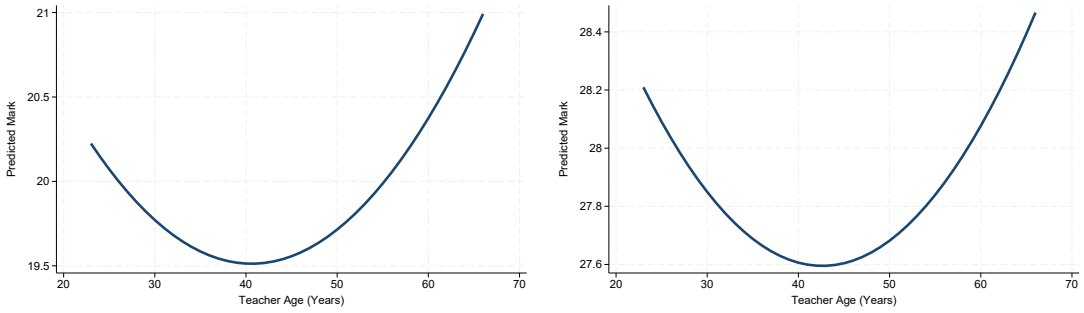
and their level of altruism. Teachers' age exhibits a significant and consistent relationship with leniency. The negative linear age coefficients from the ordered probit (−0.640, s.e.= 0.218) and OLS (−1.206, s.e.= 0.471) models, combined with the positive quadratic terms (0.693, s.e.= 0.225; 1.245, s.e.= 0.487) suggest a U-shaped pattern, whereby middle-aged teachers are relatively more stringent, while both younger and older teachers tend to grade more leniently. By contrast, gender, tenure, and teaching subject (STEM/not STEM) show no significant association with leniency in either specification. Figure 5 displays predicted leniency by age, based on the regression estimates from ordered probit (panel a) and OLS (panel b). The figure highlights that the *size* of the estimated differences across ages is not trivial. The probit estimates suggest that, when evaluating the same transcript, a 60 year-old teacher tends to assign a numerical grade that is 0.86 points (0.21 standard deviations), compared to a 40 year-old colleague. The corresponding OLS estimates would be, respectively, 0.47 points or 0.08 standard deviations.

Among behavioral characteristics, altruism is the only significant predictor of leniency (ordered probit: 0.058 , s.e. =0.018 ; OLS: 0.128 , s.e. =0.039), with more altruistic teachers grading significantly more leniently. A one-standard deviation increase in the amount donated in the dictator game is associated with about 0.1 standard deviations higher grades (0.057-to-0.128, depending on the estimator). Job satisfaction and self-efficacy show no significant relationship with leniency. Modern teaching style and effort-oriented mindset show only weakly positive associations, and the coefficients do not pass the False Discovery Rate correction threshold.

Table 3: Relationship Between Teacher Characteristics and Leniency

	F.E. Ord. Probit				F.E. OLS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.020			-0.027	-0.051			-0.070
	(0.035)			(0.035)	(0.077)			(0.077)
Age	**-0.510*****			**-0.640*****	**-0.918*****			**-1.206****
	(0.160)			**(0.218)**	**(0.347)**			**(0.471)**
Age squared	**0.534*****			**0.693*****	**0.965*****			**1.245****
	(0.161)			**(0.225)**	**(0.350)**			**(0.487)**
STEM teacher		-0.034		-0.027		-0.076		-0.069
		(0.036)		(0.036)		(0.079)		(0.079)
Tenure		-0.111*		0.071		-0.142		0.221
		(0.066)		(0.087)		(0.143)		(0.190)
Tenure squared		0.121*		-0.103		0.190		-0.203
		(0.067)		(0.095)		(0.147)		(0.208)
Job satisfaction			0.019	0.017			0.062	0.056
			(0.018)	(0.018)			(0.040)	(0.040)
Self-efficacy			-0.023	-0.021			-0.038	-0.037
			(0.019)	(0.019)			(0.041)	(0.041)
Altruism			**0.057*****	**0.058*****			**0.127*****	**0.128*****
			(0.018)	**(0.018)**			**(0.039)**	**(0.039)**
Effort–ability mindset			0.032*	0.032*			0.088**	0.086**
			(0.018)	(0.018)			(0.039)	(0.039)
Modern–trad. pedagogy			0.033*	0.034*			0.063	0.069*
			(0.019)	(0.019)			(0.040)	(0.039)
Observations	12368	12368	12368	12368	16440	16440	16440	16440
Transcript FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Response time	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each group of four columns shows standardised regressions with transcript fixed effects. Columns 1-4 use ordered probit (excluding the two highest- and two lowest-performing transcripts); columns 5-8 use OLS on the full transcript sample. The dependent variable is numerical grade (0-30). Pseudo R^2 = 0.16 in columns (1)-(4); R^2 = 0.81 in columns (5)-(8). Standard errors are clustered at the teacher level and shown in parentheses. Significance stars reflect p-values: * p<0.10, ** p<0.05, *** p<0.01. We apply the Benjamini-Hochberg (1995) procedure to control the False Discovery Rate at q=0.05 across all 44 coefficient tests in the table. Surviving significant coefficients are in bold.



(a) Ordered Probit (b) OLS

Figure 5: Grading style as a function of age

Reading note: The panels show predicted numerical grades as a function of teacher age. Panel (a) displays ordered probit estimates for the reduced sample (excluding the two highest performing transcripts). Panel (b) displays OLS estimates for the full sample. All estimates are based on full-regression models including professional and behavioral characteristics as covariates, with transcript fixed effects and response time controls. All control variables are held at their means. N=12,368 in the ordered probit model. N=16,440 in the OLS model. The means and y-axes differ between the two figures, because Ordered Probit regressions are estimated on a sample that excludes the highest performing transcripts.

Result 3a: Leniency displays a U-shape relationship with age, whereby middle-aged teachers are less likely to assign high grades.
Result 3b: Altruism is positively associated with leniency.

4.4 The predictors of *perceived* leniency

Although most teachers think to be more lenient than their average colleague (result 2a), perceived leniency varies across teachers. In this section, we examine whether the characteristics associated with actual leniency also shape teachers' self-perceptions. Regression coefficients are corrected for False Discovery Rate (q<0.05) and surviving significant coefficients are reported in bold.

Table 4 reports the results, which can be organized in three groups. First, those who are relatively more/less lenient and seem to be aware of it. This is the case for age. Indeed, age displays a U-shaped relationship with perceived leniency, with perceptions of relative leniency lowest around age 40 (−5.157, s.e.= 1.988; 5.730, s.e.= 2.087). This mirrors the U-shaped relationship found in the grading task, indicating that age-related self-perceptions are relatively well-calibrated. Second, some teachers are relatively more lenient in practice, but do not seem to be aware of it. This is the case of altruistic individuals. Those who donate more in the dictator game do not report higher perceived leniency compared to other groups, although they actually assign relatively higher grades. This result suggests that altruist teachers grade more generously without being aware of it. Finally, the third group is formed of those who believe they are particularly lenient, but for whom this is not observed in the data. Among demographic characteristics, women perceive themselves as relatively more lenient than men (0.769, s.e.= 0.302), a finding that contrasts with the absence of a gender effect on actual leniency. Among behavioral characteristics, teachers with a relatively more effort-oriented mindset (0.432, s.e.= 0.160) and those who adopt a more modern teaching style (0.813, s.e.= 0.169) perceive themselves as significantly more lenient too, but this does not match with their behavior in the grading task.

In sum, teachers' beliefs about their own leniency are only partially aligned with their actual grading behavior. We observe systematic miscalibration for some groups, but our experiment caution against using self-assessments as evidence of differential leniency.

Result 4a: Perceived leniency displays a U-shape relationship with age, thus suggesting that teachers are, to some extent, aware of their grading styles.
Result 4b: Gender and self-reported teaching styles are associated with perceived leniency, although we observe no difference in actual grading behavior.

Table 4: Relationship Between Teacher Characteristics and Perceived Leniency

	(1)	(2)	(3)	(4)
Female	**0.743****			**0.769****
	(0.310)			**(0.302)**
Age	**-3.712*****			**-5.157*****
	(1.403)			**(1.988)**
Age squared	**3.639****			**5.730*****
	(1.422)			**(2.087)**
STEM teacher		-0.019		0.042
		(0.322)		(0.318)
Tenure		-0.842		0.477
		(0.601)		(0.828)
Tenure squared		0.537		-1.182
		(0.618)		(0.922)
Job satisfaction			0.114	0.142
			(0.170)	(0.171)
Self-efficacy			0.166	0.226
			(0.169)	(0.170)
Altruism			0.215	0.230
			(0.158)	(0.156)
Effort–ability mindset			**0.438*****	**0.432*****
			(0.159)	**(0.160)**
Modern–trad. pedagogy			**0.878*****	**0.813*****
			(0.169)	**(0.169)**
Observations	1644	1644	1644	1644
Response time	Yes	Yes	Yes	Yes
R-squared	0.009	0.004	0.030	0.041

Notes: Standardised OLS regressions on the teacher-level sample (N=1,644). The dependent variable is perceived leniency. Standard errors are clustered at the teacher level and shown in parentheses. Significance stars reflect raw p-values: * p<0.10, ** p<0.05, *** p<0.01. We apply the Benjamini-Hochberg (1995) procedure to control the False Discovery Rate at q=0.05 across all 22 coefficient tests in the table. Surviving significant coefficients are in bold.

5 Conclusion

While grading discrepancies are well documented, the literature has largely focused on discriminatory biases and ingroup favoritism, leaving teachers' leniency relatively unexplored. Our experimental setting allows to measure individual leniency neatly, and study the demographic, professional and behavioral characteristics associated with it. We show that differential leniency explains a large portion of why different teachers who evaluate the same student give different grades. When comparing actual grading with the relevant beliefs, we document that teachers are often miscalibrated in the assessment of their own leniency, but that the amount of variance is within what they find acceptable, contrary to what the literature has found for other occupations.

We identify two main predictors of teachers' leniency. First, leniency is positively associated with altruism: teachers who donate more in a real-stakes dictator game tend to assign higher grades. Whether this reflects a stable personality trait, a shared norm of benevolence toward those being evaluated, or a common belief about the consequences of withholding resources from others is a question our design cannot resolve, but the cross-domain consistency echoes broader evidence on the generality of prosocial preferences (Borghans et al., 2008). Second, leniency follows a U-shaped relationship with age, reaching a minimum in the early forties, when teachers tend to be strictest. The U-shaped age profile has potential implications for the composition of grading committees. To illustrate, consider committee A, composed of two 40-year-old teachers, and committee B, where both teachers are at least 15 years above/below 40. Our estimates imply that, for two students of identical quality, the one evaluated by committee A would receive, on average, a grade approximately 1.7 percentage points lower than the other (i.e., 0.5 points on a 0–30 scale).

This example highlights how differential leniency can affect grading outcomes in a way that is, potentially, preventable. Our results also provide empirical ground for interventions aimed at debiasing and promoting a fair assessment system. We hope future research will seek to replicate and extend these findings in other settings, with the medium-term aim of developing concrete guidelines for educational institutions.

References

Alan, S., Ertac, S., and Mumcu, I. (2018). Gender stereotypes in the classroom and effects on achievement. *Review of Economics and Statistics*, 100(5):876–890.

Alesina, A., Carlana, M., Ferrara, E. L., and Pinotti, P. (2024). Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review*, 114(7):1916–1948.

Backes, B., Cowan, J., Goldhaber, D., and Theobald, R. (2024). How to measure a teacher: The influence of test and nontest value-added on long-run student outcomes. *Journal of Human Resources*.

Bagues, M. and Perez-Villadoniga, M. J. (2012). Do recruiters prefer applicants with similar skills? evidence from a randomized natural experiment. *Journal of Economic Behavior & Organization*, 82(1):12–20.

Bagues, M., Sylos-Labini, M., and Zinovyeva, N. (2017). Does the gender composition of scientific committees matter? *American Economic Review*, 107(4):1207–1238.

Bandiera, O., Barankay, I., and Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics*, 120(3):917–962.

Bandura, A. and Adams, N. E. (1977). Analysis of self-efficacy theory of behavioral change. *Cognitive therapy and research*, 1(4):287–310.

Bénabou, R. (2015). The economics of motivated beliefs. *Revue d'économie politique*, 125(5):665–685.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.

Benjamin, D. J., Cooper, K., Heffetz, O., Kimball, M. S., and Zhou, J. (2023). Adjusting for scale-use heterogeneity in self-reported well-being. Technical report, National Bureau of Economic Research.

Borghans, L., Duckworth, A. L., Heckman, J. J., and Weel, B. t. (2008). The economics and psychology of personality traits. *Journal of human Resources*, 43(4):972–1059.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., and Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management science*, 62(10):2765–2783.

Bredtmann, J., Otten, S., and Vonnahme, C. (2024). Discrimination in grading? evidence on teachers' evaluation bias towards minority students. (1122).

Burgess, S. and Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31(3):535–576.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

Carlana, M., La Ferrara, E., and Pinotti, P. (2022). Goals and gaps: Educational careers of immigrant children. *Econometrica*, 90(1):1–29.

Carrell, S. E., Page, M. E., and West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *The Quarterly journal of economics*, 125(3):1101–1144.

Carrell, S. E. and West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432.

Charness, G. and Gneezy, U. (2008). What's in a name? anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization*, 68(1):29–35.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9):2593–2632.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources*, 45(3):655–681.

Contreras, D. (2024). Gender differences in grading: Teacher bias or student behaviour? *Education Economics*, 32(6):762–785.

Copur-Gencturk, Y., Thacker, I., and Quinn, D. (2021). K-8 teachers' overall and gender-specific beliefs about mathematical aptitude. *International Journal of Science and Mathematics Education*, 19(6):1251–1269.

Currie, J. and MacLeod, W. B. (2017). Diagnosing expertise: Human capital, decision making, and performance among physicians. *Journal of labor economics*, 35(1):1–43.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

de Gendre, A., Feld, J., Salamanca, N., and Zölitz, U. (2024). Same-sex teacher effects. Technical report, Working Paper.

De Neve, J.-E. and Ward, G. (2025). *Why Workplace Wellbeing Matters: The Science Behind Employee Happiness and Organizational Performance*. Harvard Business Press.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of economics and statistics*, 86(1):195–210.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human resources*, 42(3):528–554.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2022). Estimating social preferences and gift exchange at work. *American Economic Review*, 112(3):1038–1074.

Denning, J. T., Nesbit, R. L., Pope, N. G., and Warnick, M. (2026). Easy a's, less pay: The long-term effects of grade inflation. Technical report, National Bureau of Economic Research.

Deschamps, P. (2023). Gender quotas in hiring committees: A boon or a bane for women? *Management Science*, 70(11):7486–7505.

Dhawan, V., Bramley, T., and Assessment, C. (2012). Estimation of inter-rater reliability. *Coventry, UK: Ofqual*.

Diris, R., Isphording, I. E., Landaud, F., and Sievertsen, H. H. (2026). Grades and assessments – the economists' perspective. In Bradley, S. and Green, C., editors, *The Economics of Education*. 3 edition. Forthcoming, version March 16, 2026.

Dobbie, W., Goldin, J., and Yang, C. S. (2018). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–240.

Doyle, L., Easterbrook, M. J., and Harris, P. R. (2023). Roles of socioeconomic status, ethnicity and teacher beliefs in academic grading. *British Journal of Educational Psychology*, 93(1):91–112.

Dur, R. and Van Lent, M. (2018). Serving the public interest in several ways: Theory and empirics. *Labour Economics*, 51:13–24.

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings–part i. *Psychological Test and Assessment Modeling*, 59(4):443–452.

Engel, E., Hayes, R. M., and Wang, X. (2010). Audit committee compensation and the demand for monitoring of the financial reporting process. *Journal of accounting and economics*, 49(1-2):136–154.

Epley, N. and Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic perspectives*, 30(3):133–140.

Feld, J., Salamanca, N., and Hamermesh, D. S. (2016). Endophilia or exophobia: Beyond discrimination. *The Economic Journal*, 126(594):1503–1527.

Filetti, J., Wright, M., and King, W. M. (2010). Grades and ranking: When tenure affects assessment. *Practical Assessment, Research & Evaluation*, 15(1):14.

Gershenson, S., Hart, C. M., Hyman, J., Lindsay, C. A., and Papageorge, N. W. (2022). The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy*, 14(4):300–342.

Gershenson, S., Holt, S. B., and Papageorge, N. W. (2016). Who believes in me? the effect of student–

teacher demographic match on teacher expectations. *Economics of education review*, 52:209–224.

Gong, J., Lu, Y., and Song, H. (2018). The effect of teacher gender on students' academic and noncognitive outcomes. *Journal of Labor Economics*, 36(3):743–778.

Griffith, A. L. and Sovero, V. (2021). Under pressure: How faculty gender and contract uncertainty impact students' grades. *Economics of Education Review*, 83:102126.

Hoffman, M., Kahn, L. B., and Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale esl writing assessment. *Assessing Writing*, 17(3):123–139.

Huang, J. and Whipple, P. B. (2023). Rater variability and reliability of constructed response questions in new york state high-stakes tests of english language arts and mathematics: implications for educational assessment policy. *Humanities and Social Sciences Communications*, 10(1):1–10.

Huang, J., Zhu, D., Xie, D., and Shu, T. (2023). Examining the reliability of an international chinese proficiency standardized writing assessment: Implications for assessment policy makers. *Assessing Writing*, 55:100693.

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Hachette UK.

Kaiser, C. (2022). Using memories to assess the intrapersonal comparability of wellbeing reports. *Journal of Economic Behavior & Organization*, 193:410–442.

Kates, S., Paulsen, T., Yntiso, S., and Tucker, J. A. (2023). Bridging the grade gap: Reducing assessment bias in a multi-grader class. *Political Analysis*, 31(4):642–650.

Kezim, B., Pariseau, S. E., and Quinn, F. (2005). Is grade inflation related to faculty status? *Journal of Education for Business*, 80(6):358–364.

King, G., Murray, C. J., Salomon, J. A., and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, 98(1):191–207.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Klimecki, O. M., Mayer, S. V., Jusyte, A., Scheeff, J., and Schönenberg, M. (2016). Empathy promotes altruistic behavior in economic interactions. *Scientific reports*, 6(1):31961.

Kline, P., Rose, E. K., and Walters, C. R. (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, 137(4):1963–2036.

Kosse, F. and Tincani, M. M. (2020). Prosociality predicts labor market success around the world. *Nature communications*, 11(1):5298.

Kristoffersen, J. H. G., Krægpøth, M. V., Nielsen, H. S., and Simonsen, M. (2015). Disruptive school peers and student outcomes. *Economics of Education Review*, 45:1–13.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Public Economics*, 92(10-11):2083–2105.

Lavy, V. and Megalokonomou, R. (2019). Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study. Working Paper 26021, National Bureau of Economic Research.

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

Lavy, V., Sand, E., and Shayo, M. (2022). Discrimination between religious and non-religious groups: Evidence from marking high-stakes exams. *The Economic Journal*, 132(646):2308–2324.

Lavy, V. and Yancu, A. (2025). Violent peers at school: Impacts and mechanisms. Technical report, National Bureau of Economic Research.

Lim, J. and Meer, J. (2017). The impact of teacher–student gender matches: Random assignment evidence from south korea. *Journal of Human Resources*, 52(4):979–997.

Lusher, L., Campbell, D., and Carrell, S. (2018). Tas like me: Racial interactions between graduate

teaching assistants and undergraduates. *Journal of Public Economics*, 159:203–224.

MENJS (2021). Bilan social 2020-2021. *Rapport*.

Monnet, M., Colo, P., and Dagorn, E. (2025). Determinants of gender discrimination by teachers: Evidence from an online experiment. *Available at SSRN 5254180*.

Myford, C. M. and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part i. *Journal of applied measurement*, 4(4):386–422.

Oswald, A. J. (2008). On the curvature of the reporting function from objective reality to subjective feelings. *Economics Letters*, 100(3):369–372.

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., Ford, C. E., and Carnes, M. (2018). Low agreement among reviewers evaluating the same nih grant applications. *Proceedings of the National Academy of Sciences*, 115(12):2952–2957.

Protivínskỳ, T. and Münich, D. (2018). Gender bias in teachers' grading: What is in the grade. *Studies in Educational Evaluation*, 59:141–149.

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, 42(3):375–392.

Sahlström, E. and Silliman, M. (2024). The extent and consequences of teacher biases against immigrants.

Seré, M. and Decancq, K. (2025). Room for happiness? cross-country variation in response scale. Mimeo.

Taylor, E. S. and Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7):3628–3651.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, 77:101981.

U.S. Sentencing Commission (1991). The federal sentencing guidelines: A report on the operation of the guidelines system and short-term impacts on disparity in sentencing, use of incarceration, and prosecutorial discretion and plea bargaining. Technical report, U.S. Sentencing Commission, Washington, DC. Volumes 1 & 2.

Winters, M. A., Haight, R. C., Swaim, T. T., and Pickering, K. A. (2013). The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review*, 34:69–75.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, 100:61–78.

Zell, E., Strickhouser, J. E., Sedikides, C., and Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological bulletin*, 146(2):118.

Zhu, M. (2024). New findings on racial bias in teachers' evaluations of student achievement. Discussion Paper 16815, Institute of Labor Economics (IZA).

Acknowledgments

We would like to thank Philippe Colo and Marion Monnet, without whom this study would have never been possible. Our thanks go also to Gustave Kenedi, Victor Lavy and Mikko Silliman for their precious comments and suggestions; to the staff from the local authority of education for their essential support; as well as to Yann Venance and Max R. P. Grossmann for their technical help. We also thank participants in the 2026 London Behavioral and Experimental Economics workshop and the 2026 ASFEE conference. Financial support from the Women and Science Chair and PSE is gratefully acknowledged. The French State under the France-2030 programme

and the Initiative of Excellence of the University of Lille are acknowledged for the funding and support granted to the R-CDP-24-006-DePERU project. The main hypotheses were pre-registered in the OSF Registry: https://osf.io/nphvk. The research and procedure were approved by the ETH Zurich Ethics Commission, number EK 2023-N-10, and the INED Data Protection Office (INED 2023-DPD-0009). All errors are our own.

Online Appendix

Teachers' Leniency:

Experimental Evidence

A Appendix 1: Additional figures and tables

A.1 Tables

Table ST-1: Descriptive Statistics of Teacher Characteristics

	Mean	Std. Dev.	Min	Max
Female	0.64	0.48	0	1
Age (years)	44.92	8.87	23	66
Tenure (years)	18.17	9.21	0	42
STEM teacher	0.41	0.49	0	1
Job satisfaction (0-10)	7.05	1.55	0	10
Self confidence (1-7)	3.57	1.58	1	7
Perceived Leniency	2.13	6.57	-24	30
Acceptable Deviation	4.87	3.94	0	30
Altruism	28.97	14.33	0	50
Modern–trad. pedagogy	4.32	0.77	1	7
Effort–ability mindset	4.74	1.09	1	7
Response time (seconds)	46.35	18.24	21	225

Notes: N = 1,644

Table ST-2: Fictitious School Transcript Characteristics

Profile number	Gender		Science Performance		Humanities Performance		Behaviour in class	
	Male	Female	Low	High	Low	High	Poor	Good
Profile 1		✓		✓		✓		✓
Profile 2		✓		✓		✓	✓	
Profile 3		✓		✓	✓			✓
Profile 4		✓		✓	✓		✓	
Profile 5		✓	✓			✓		✓
Profile 6		✓	✓			✓	✓	
Profile 7		✓	✓		✓			✓
Profile 8		✓	✓		✓		✓	
Profile 9	✓			✓		✓		✓
Profile 10	✓			✓		✓	✓	
Profile 11	✓			✓	✓			✓
Profile 12	✓			✓	✓		✓	
Profile 13	✓		✓			✓		✓
Profile 14	✓		✓			✓	✓	
Profile 15	✓		✓		✓			✓
Profile 16	✓		✓		✓		✓	

Notes: The table presents the 16 fictitious transcripts characteristics. The 16 profiles are built varying gender, performance in science (low or high) that is determined based on grades in maths and biology, performance in humanities (low or high) that is determined based on grades in French, English and history, and behaviour (poor or good) based on the number of late arrivals and unjustified absences. Figure 1, in the main text, corresponds to Profile 9.

Table ST-3: Grades, leniency and noise by predictor

(1) Variable	(2) Split	(3) Grade	(4) Benchmark	(5) Leniency	(6) Noise	(7) Perc. Leniency	(8) Accept. Deviation	(9) N
Gender	Male	19.73	19.70	0.03[-0.00, 0.07]	1.57[1.55, 1.59]	1.67[1.54, 1.81]	4.53[4.44, 4.61]	6000
	Female	19.75	19.77	-0.02[-0.05, 0.01]	1.75[1.74, 1.77]	2.40[2.26, 2.53]	5.07[4.99, 5.15]	10440
STEM teacher	No	19.79	19.76	0.03[0.00, 0.06]	1.70[1.69, 1.72]	2.13[1.99, 2.27]	4.99[4.90, 5.08]	9720
	Yes	19.67	19.72	-0.05[-0.08, -0.01]	1.66[1.64, 1.68]	2.14[2.00, 2.28]	4.71[4.63, 4.78]	6720
Age	Low	19.68	19.76	-0.08[-0.11, -0.05]	1.68[1.66, 1.70]	2.18[2.05, 2.31]	4.77[4.70, 4.85]	8950
	High	19.81	19.72	0.09[0.06, 0.13]	1.70[1.68, 1.72]	2.07[1.92, 2.23]	5.00[4.90, 5.09]	7490
Tenure	Low	19.72	19.76	-0.04[-0.07, -0.01]	1.72[1.70, 1.74]	2.49[2.35, 2.62]	4.90[4.81, 4.98]	8340
	High	19.76	19.72	0.04[0.01, 0.08]	1.65[1.64, 1.67]	1.77[1.62, 1.91]	4.85[4.77, 4.94]	8100
Job satisfaction	Low	19.69	19.73	-0.05[-0.08, -0.02]	1.71[1.69, 1.72]	2.11[1.96, 2.25]	5.07[4.98, 5.15]	9160
	High	19.81	19.75	0.06[0.02, 0.10]	1.66[1.64, 1.68]	2.16[2.03, 2.30]	4.63[4.54, 4.72]	7280
Self-efficacy	Low	19.76	19.73	0.03[-0.00, 0.05]	1.65[1.64, 1.67]	2.01[1.89, 2.13]	4.88[4.81, 4.95]	11920
	High	19.70	19.77	-0.07[-0.12, -0.02]	1.77[1.75, 1.80]	2.46[2.26, 2.65]	4.86[4.75, 4.97]	4520
Altruism	Low	19.62	19.73	-0.12[-0.15, -0.08]	1.74[1.73, 1.76]	2.10[1.96, 2.25]	4.78[4.69, 4.86]	8330
	High	19.87	19.75	0.12[0.09, 0.15]	1.63[1.61, 1.65]	2.16[2.03, 2.30]	4.97[4.89, 5.06]	8110
Effort–ability	Low	19.68	19.73	-0.05[-0.08, -0.02]	1.68[1.66, 1.69]	1.88[1.76, 1.99]	4.82[4.75, 4.89]	12330
mindset	High	19.91	19.77	0.14[0.10, 0.19]	1.71[1.69, 1.74]	2.90[2.70, 3.10]	5.03[4.91, 5.16]	4110
Modern–trad.	Low	19.67	19.71	-0.04[-0.07, -0.01]	1.69[1.68, 1.71]	1.51[1.38, 1.64]	4.83[4.75, 4.91]	9780
pedagogy	High	19.84	19.78	0.06[0.02, 0.10]	1.68[1.66, 1.70]	3.05[2.89, 3.21]	4.94[4.84, 5.04]	6660
All		19.74	19.74	0.00[-0.02, 0.02]	1.69[1.67, 1.70]	2.13[2.03, 2.23]	4.87[4.81, 4.93]	16440

Notes: This table reports descriptive statistics for numerical grades (0-30 scale) decomposed by teacher characteristics. Variables are split at the median unless binary. Column (3): mean grade assigned. Column (4): benchmark grade (mean grade given by all teachers to the same transcripts). Column (5): leniency (bias), defined as the difference between columns (3) and (4). Column (6): noise, defined as the RMSE of within-teacher residuals after removing bias. Column (7): mean self-reported perceived leniency (−30 to +30). Column (8): mean self-reported acceptable deviation. Column (9): number of transcript-level observations. 95% confidence intervals in brackets. N=16,440 transcript-level observations from 1,644 teachers.

Table ST-4: Robustness test: include the anonymised transcript used in Monnet et al. (2025)

	(1) F.E. Ord. Probit	(2) F.E. OLS	(3) Simple OLS
Dep.var.	*Grade*	*Grade*	*Perceived leniency*
Female	-0.020	-0.057	**0.714****
	(0.033)	(0.073)	**(0.286)**
Age	**-0.629*****	**-1.221*****	**-5.773*****
	(0.205)	**(0.450)**	**(1.892)**
Age squared	**0.676*****	**1.256*****	**6.402*****
	(0.211)	**(0.466)**	**(1.977)**
STEM teacher	-0.024	-0.059	0.065
	(0.034)	(0.075)	(0.301)
Tenure	0.059	0.199	0.757
	(0.082)	(0.181)	(0.792)
Tenure squared	-0.082	-0.170	-1.510*
	(0.089)	(0.199)	(0.871)
Job satisfaction	0.016	0.054	0.120
	(0.017)	(0.038)	(0.160)
Self-efficacy	-0.020	-0.038	0.186
	(0.017)	(0.038)	(0.160)
Altruism	**0.055*****	**0.123*****	0.239
	(0.017)	**(0.037)**	(0.147)
Effort–ability mindset	0.021	0.066*	**0.382****
	(0.017)	(0.037)	**(0.148)**
Modern–trad. pedagogy	**0.043****	**0.085****	**0.791*****
	(0.018)	**(0.038)**	**(0.160)**
N	13872	17944	1832
Transcript FE	Yes	Yes	No
Response time	Yes	Yes	Yes
R-squared		0.808	0.040
Pseudo R-squared	0.182		

Notes: Column 1 shows standardised ordered probit regressions (excluding the two highest- and two lowest-performing transcripts). Column 2 shows standardised OLS regressions for all transcripts. The dependent variable in columns 1 and 2 is numerical grade. Standard errors are clustered at the teacher level and shown in parentheses. Column 3 shows a standardised OLS regression at the teacher level; the dependent variable is perceived leniency. All models include the full set of teacher characteristics, transcript fixed effects (except column 3), and response time controls. Significance stars reflect raw p-values: * p<0.10, ** p<0.05, *** p<0.01. We apply the Benjamini-Hochberg (1995) procedure to control the False Discovery Rate at q=0.05 across all 33 coefficient tests in the table. Surviving significant coefficients are in bold.

Table ST-5: Robustness test: Exclude teachers who display confusion about the grading scale

	(1) F.E. Ord. Probit	(2) F.E. OLS	(3) Simple OLS
Dep.var.	*Grade*	*Grade*	*Perceived leniency*
Female	-0.030	-0.073	**0.854*****
	(0.035)	(0.078)	**(0.305)**
Age	**-0.655*****	**-1.248*****	**-5.897*****
	(0.213)	**(0.466)**	**(2.012)**
Age squared	**0.689*****	**1.257*****	**6.507*****
	(0.218)	**(0.479)**	**(2.120)**
STEM teacher	-0.027	-0.075	0.013
	(0.036)	(0.079)	(0.320)
Tenure	0.093	0.258	0.595
	(0.085)	(0.189)	(0.835)
Tenure squared	-0.107	-0.211	-1.279
	(0.091)	(0.203)	(0.933)
Job satisfaction	0.015	0.049	0.091
	(0.018)	(0.041)	(0.172)
Self-efficacy	-0.023	-0.046	0.207
	(0.019)	(0.041)	(0.171)
Altruism	**0.058*****	**0.129*****	0.277*
	(0.018)	**(0.039)**	(0.158)
Effort–ability mindset	0.034*	0.090**	**0.455*****
	(0.018)	(0.040)	**(0.161)**
Modern–trad. pedagogy	0.042**	0.082**	**0.785*****
	(0.019)	(0.040)	**(0.170)**
N	12081	16080	1608
Transcript FE	Yes	Yes	No
Response time	Yes	Yes	Yes
R-squared		0.811	0.043
Pseudo R-squared	0.162		

Notes: The sample excludes teachers who used the 0-20 grade scale rather than the 0-30 scale specified in the instructions. Column 1 shows standardised ordered probit regressions (excluding the two highest- and two lowest-performing transcripts). Column 2 shows standardised OLS regressions for all transcripts. The dependent variable in columns 1 and 2 is numerical grade. Standard errors are clustered at the teacher level and shown in parentheses. Column 3 shows a standardised OLS regression at the teacher level; the dependent variable is perceived leniency. All models include the full set of teacher characteristics, transcript fixed effects (except column 3), and response time controls. Significance stars reflect raw p-values: * p<0.10, ** p<0.05, *** p<0.01. We apply the Benjamini-Hochberg (1995) procedure to control the False Discovery Rate at q=0.05 across all 33 coefficient tests in the table. Surviving significant coefficients are in bold.

Table ST-6: Relationship Between Teacher Characteristics and Mean Absolute Deviation

	(1) F.E. OLS	(2) Simple OLS
Dep.var.	*Deviation*	*Acc. deviation*
Female	**0.191*****	**0.561*****
	(0.042)	**(0.190)**
Age	-0.228	-2.841**
	(0.256)	(1.407)
Age squared	0.236	3.170**
	(0.256)	(1.478)
STEM teacher	-0.025	-0.242
	(0.042)	(0.180)
Tenure	-0.135	0.567
	(0.116)	(0.516)
Tenure squared	0.112	-0.803
	(0.117)	(0.562)
Job satisfaction	-0.025	-0.202*
	(0.022)	(0.105)
Self-efficacy	0.028	0.074
	(0.022)	(0.103)
Altruism	-0.052**	0.181*
	(0.022)	(0.094)
Effort–ability mindset	0.023	0.184*
	(0.021)	(0.096)
Modern–trad. pedagogy	-0.027	0.078
	(0.024)	(0.104)
N	16440	1644
Transcript FE	Yes	No
Response time	Yes	Yes
R-squared	0.024	0.024

Notes: The dependent variable in column 1 is per-teacher Mean Absolute Deviation (MAD): the absolute difference between a teacher's grade and the transcript-level mean. Column 1 reports standardised OLS estimates on the full sample of transcripts. Column 2 reports a standardised OLS regression at the teacher level; the dependent variable is Acceptable Deviation (the deviation each teacher reports as acceptable). Standard errors are clustered at the teacher level and shown in parentheses. Significance stars reflect raw p-values: * p<0.10, ** p<0.05, *** p<0.01. We apply the Benjamini-Hochberg (1995) procedure to control the False Discovery Rate at q=0.05 across all 22 coefficient tests in the table. Surviving significant coefficients are in bold.

A.2 Figures

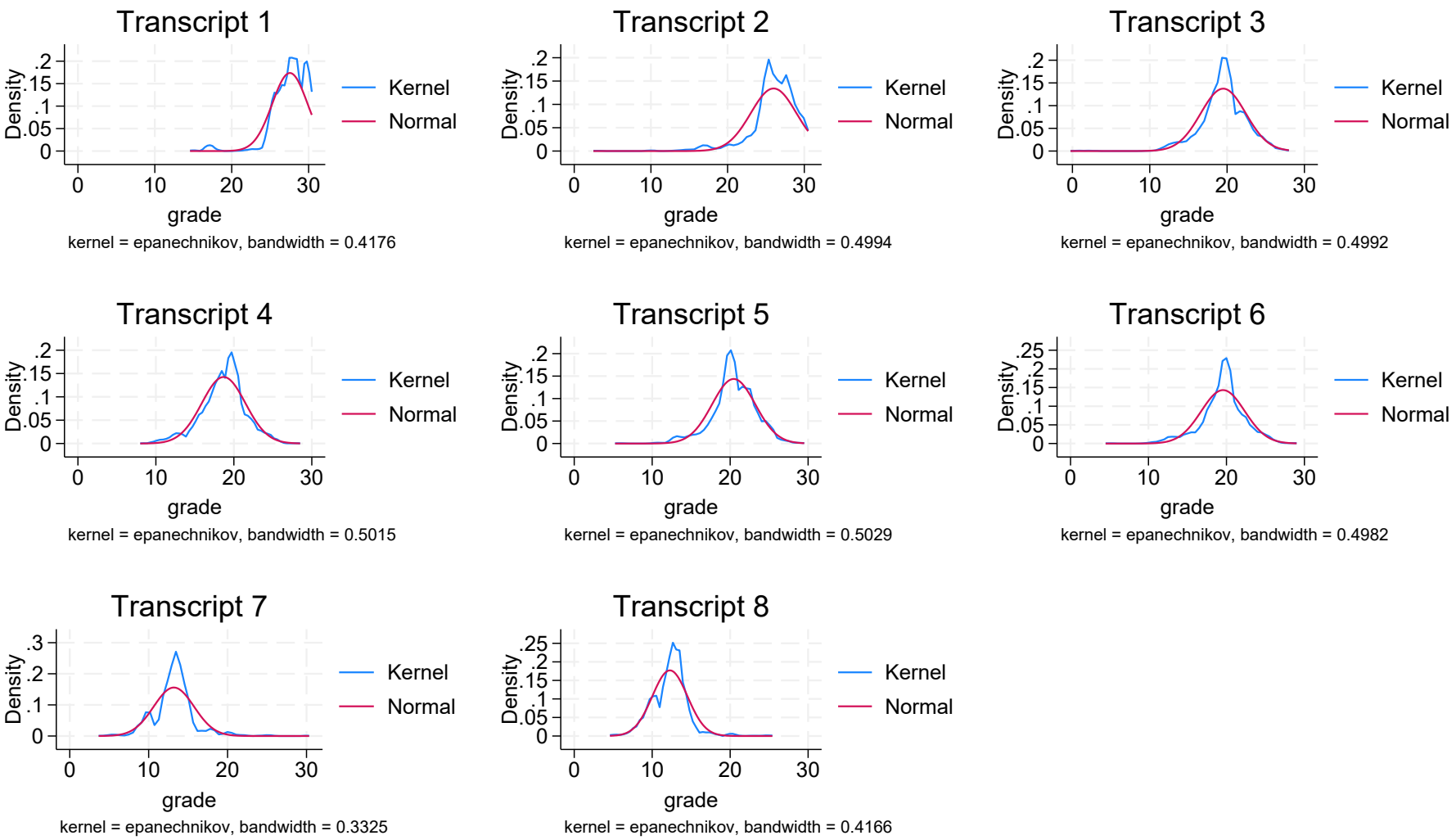


Figure SF-1: Distribution of female transcripts

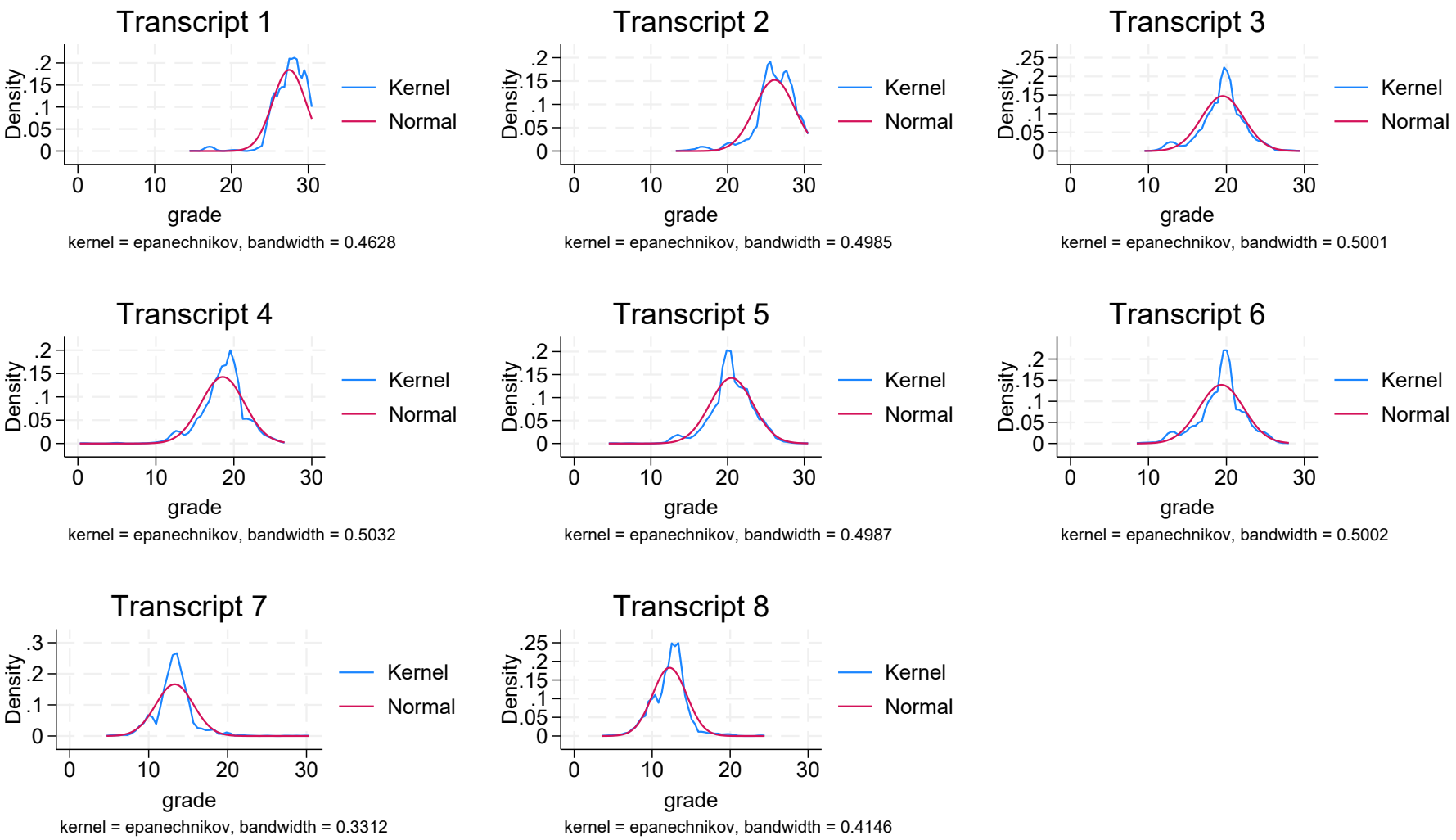
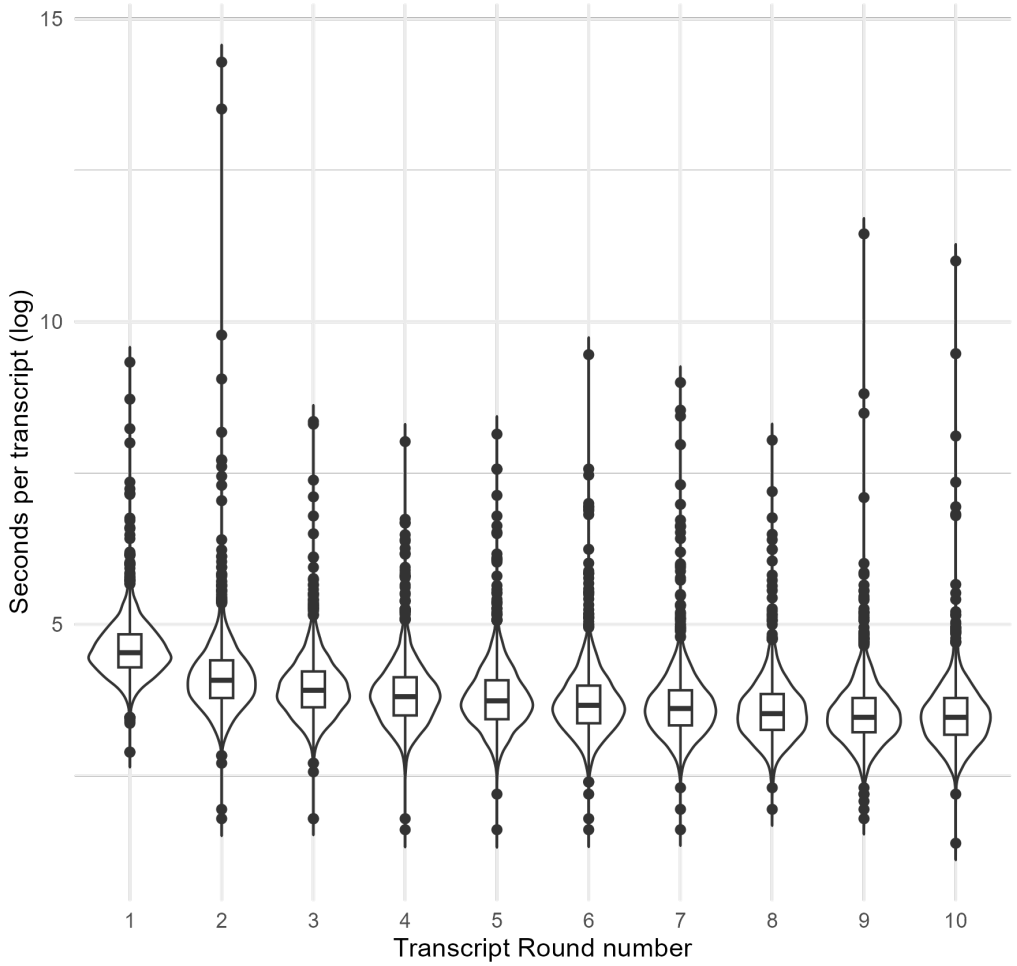


Figure SF-2: Distribution of male transcripts

Figure SF-3: Distribution of the (log) Time Spent on Each Transcript by Round Number



Notes: This figure shows the distribution of time (log-transformed seconds) spent on each transcript page across the 10 different rounds. The x-axis represents the round number of the transcripts, and the y-axis shows the log of the seconds spent on the page. The violin plot illustrates the density distribution for each round, while the boxplot inside provides a summary of the central tendency and variability (median and interquartile range).

B Appendix 2: Illustrative toy example

To fix the ideas, imagine three teachers (A, B, and C) evaluating five transcripts from five students (S1, S2, S3, S4, S5) on a scale from 0 to 10. Teacher A gives 5 out of 10 to S1, 6 out of 10 to S2, and so on. The table also indicates the mean grade given to each transcript and the mean grade given by each teacher. Looking at the last row, variations in grades for the same transcript indicate the combination of leniency (l_t) and noise (n_{it}). Looking at the last column, differences in average grades among teachers identifies differences in leniency only (l_t). Indeed, it is apparent that teacher A is relatively more lenient, while teacher C is relatively harsher.

	Transcr. 1	**Transc. 2**	**Transcr. 3**	**Transcr. 4**	**Transcr. 5**	***Mean***
Teacher A	5	6	7	8	9	**7.00**
Teacher B	5	6	7	8.5	8	**6.90**
Teacher C	4	6	8	6	8	**6.40**
Mean	**4.67**	**6.00**	**7.33**	**7.50**	**8.33**	**6.77**

Table ST-7: Illustrative example with 3 teachers and 5 students

To measure the amount of variability due to differential leniency, we calculate the squared mean error of teacher's grades, i.e.:

$$
\begin{aligned}
\text{Leniency} &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{S}\sum_{s=1}^{S}(x_{st}-\bar{x}_s)\right]^2 \\
&= \frac{1}{3}\sum_{t=1}^{3}\left[(x_{1t}-\bar{x}_1)+(x_{2t}-\bar{x}_2)+(x_{3t}-\bar{x}_3)+(x_{4t}-\bar{x}_4)+(x_{5t}-\bar{x}_5)\right]^2 \\
&= \frac{1}{3}\left[(ME_A)^2+(ME_B)^2+(ME_C)^2\right] \\
&= \frac{1}{3}\left[(0.23)^2+(0.13)^2+(-0.37)^2\right] \\
&= 0.07
\end{aligned}
$$

Our individual measure of leniency corresponds to the Mean Error (ME) of each teacher, i.e. the difference between their average grade and the total average grade. For instance, the leniency score is 0.23 for teacher A, calculated as: 7.00-6.77.

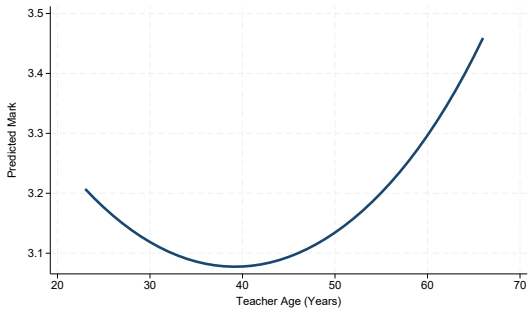
C Appendix 3: Analysis of verbal marks

The verbal mark ordered options were: “concerning”; “need more effort / need better conduct”; “encouragements”; “good/honors”; “congratulations”; “excellence/high honors”. The options “need more effort” and “need better conduct” were presented as separate options to the respondent, but we merge them in our analyses, since they do not have a clear relative order. Table ST-8 presents a parallel ANOVA for verbal marks (1-7 scale), showing a similar pattern with some notable differences. Transcript quality remains the dominant factor, accounting for 77.1% of the variance in verbal marks. Here, leniency explains about 6.8% of the variance. That is, after controlling for transcript fixed-effects, teacher leniency accounts for about 30% of the remaining variance in varbal marks. Both transcript ($F = 4731$, $p < 0.00$) and teacher ($F = 3.8$, $p < 0.001$) effects are highly significant determinants of verbal marks.

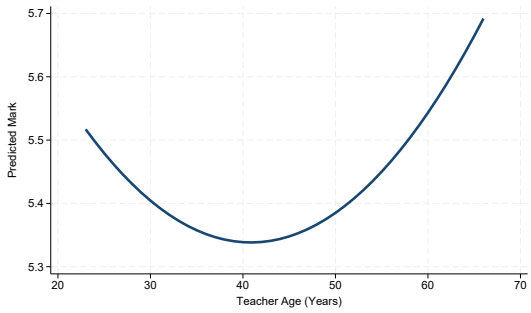
Table ST-8: Detailed Analysis of Variance Results: Effects of transcript and teacher on verbal grade.

Source of variance	Unique variance	Sum of squares	Mean squares	F-test ratio	Significance
Transcript	77.1%	22,462	1,497	4731	<0.001
Teacher	6.8%	1,980	1	3.8	<0.001
Residual (and occasion noise)	16.1%	4,678	0		

Notes: This table decomposes the total variance in transcript mentions into three components. The ‘Transcript’ component (77.1%) represents systematic differences in quality between transcripts. The ‘Teacher’ component (6.8%) captures systematic differences in mention assignment standards between teachers (level noise). The ‘Residual’ component (16.1%) represents the combined effect of teacher-transcript interactions and random occasion noise. Analysis based on 16,440 mentions from French teachers evaluating 10 student transcripts. The F-test shows that both transcript quality and teacher effects are highly significant determinants of verbal grade.



(a) Ordered Probit - Verbal



(b) OLS - Verbal

Table ST-9: Relationship Between Teacher Characteristics and Leniency

	F.E. Ord. Probit				F.E. OLS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.010			-0.003	-0.026			-0.025
	(0.039)			(0.039)	(0.018)			(0.018)
Age	**-0.772*****			**-0.819*****	**-0.314*****			**-0.399*****
	(0.142)			**(0.196)**	**(0.065)**			**(0.092)**
Age squared	**0.774*****			**0.918*****	**0.309*****			**0.428*****
	(0.143)			**(0.201)**	**(0.067)**			**(0.095)**
STEM teacher		0.033		0.042		0.010		0.010
		(0.039)		(0.039)		(0.018)		(0.018)
Tenure		**-0.259*****		-0.047		**-0.084*****		0.027
		(0.065)		(0.089)		**(0.031)**		(0.043)
Tenure squared		**0.227*****		-0.068		0.068**		-0.067
		(0.066)		(0.093)		(0.032)		(0.045)
Job satisfaction			-0.001	-0.000			0.003	0.002
			(0.019)	(0.019)			(0.009)	(0.009)
Self-efficacy			-0.001	0.006			0.001	0.004
			(0.018)	(0.019)			(0.009)	(0.009)
Altruism			0.023	0.028			0.014	0.016*
			(0.019)	(0.019)			(0.009)	(0.009)
Effort–ability mindset			0.001	-0.000			-0.001	-0.003
			(0.019)	(0.019)			(0.009)	(0.008)
Modern–trad. pedagogy			**0.068*****	**0.060*****			**0.033*****	**0.030*****
			(0.020)	**(0.020)**			**(0.009)**	**(0.009)**
Observations	12368	12368	12368	12368	16440	16440	16440	16440
Transcript FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Response time	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each group of four columns shows standardised regressions with transcript fixed effects. Columns 1-4 use ordered probit (excluding the two highest- and two lowest-performing transcripts); columns 5-8 use OLS on the full transcript sample. The dependent variable is verbal grade (1-7). Standard errors are clustered at the teacher level and shown in parentheses. Pseudo R^2 = 0.27 in columns (1)-(4); R^2 = 0.78 in columns (5)-(8). Significance stars reflect raw p-values: * p<0.10, ** p<0.05, *** p<0.01. We apply the Benjamini-Hochberg (1995) procedure to control the False Discovery Rate at q=0.05 across all 22 coefficient tests in the table. Surviving significant coefficients are in bold.

D Appendix 4: Survey questionnaire

This appendix reproduces the post-experimental questionnaire administered to participants. All items were translated from French. Unless otherwise specified, responses were collected using 7-point Likert scales. Within Block 2 and 3, items were displayed in randomized order. Conceptually similar items were included across blocks using alternative wording to improve scale reliability.

D.1 Transition

[Transition screen — no questions.]

D.2 Belief elicitation about perceived leniency

At the beginning of the questionnaire, you assigned grades from 0 to 30 to different transcripts.

In your opinion, what would be the average difference between the grade you assigned to one of these transcripts and the grade one of your colleagues would have given?

Among the transcripts you evaluated in the first task of the survey, we will randomly draw one. If you correctly guessed the difference between your grade and the average grade assigned by your colleagues (within plus or minus 0.25 points), you could be entered into a draw to receive an additional 10 euros.

[Slider from −30 (“My grade is lower”) to +30 (“My grade is higher”), with midpoint “My grade is equal”.]

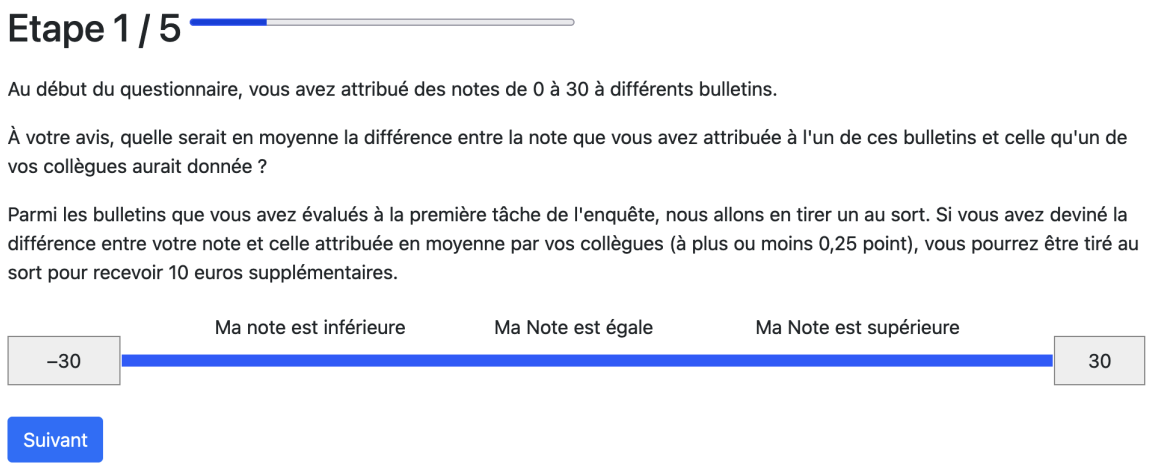


Figure SF-5: Screenshot of the belief-elicitation task (translated above).

D.3 Belief elicitation about acceptable deviation

At the beginning of the questionnaire, you assigned grades from 0 to 30 to different transcripts.

Now imagine that these transcripts are being evaluated as part of an admissions process for higher education, and that the same transcript is evaluated by two teachers.

In your opinion, what would be the difference between the two grades that you would *consider acceptable* to ensure a fair and equitable process?

[Slider from 0 to 30.]

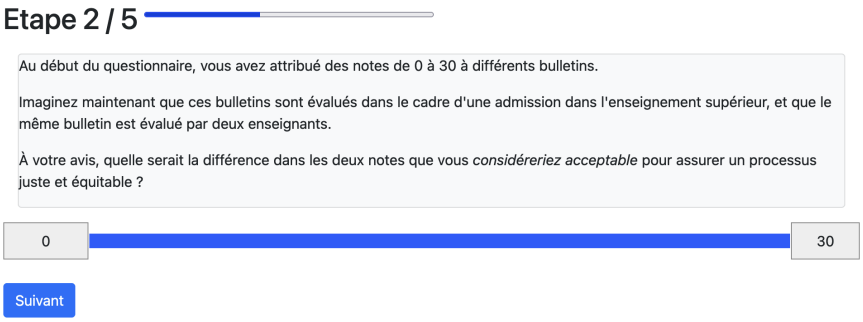


Figure SF-6: Screenshot of the fairness-perception task (translated above).

D.4 Job satisfaction

[Included in Block 1]

How satisfied are you with your job in general?

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐

D.5 Pedagogical Beliefs (Block 1)

Please indicate the extent to which you agree or disagree with each of the following statements on a scale from 1 (strongly disagree) to 7 (strongly agree).

Innate ability beliefs

S1. To be a good student, one needs particular predispositions that cannot be taught.

1 ☐ Strongly disagree — 2 ☐ — 3 ☐ — 4 ☐ Neither agree nor disagree — 5 ☐ — 6 ☐ — 7 ☐ Strongly agree

S2. Anyone can become a good student by putting in effort and staying committed.

1 ☐ Strongly disagree — 2 ☐ — 3 ☐ — 4 ☐ Neither agree nor disagree — 5 ☐ — 6 ☐ — 7 ☐ Strongly agree

Gender beliefs

S3. Even if it is not politically correct to say so, boys are often better at mathematics than girls.

1 ☐ Strongly disagree — 2 ☐ — 3 ☐ — 4 ☐ Neither agree nor disagree — 5 ☐ — 6 ☐ — 7 ☐ Strongly agree

S4. Girls often have to work harder than boys to be good at mathematics.

1 ☐ Strongly disagree — 2 ☐ — 3 ☐ — 4 ☐ Neither agree nor disagree — 5 ☐ — 6 ☐ — 7 ☐ Strongly agree

Teaching style

S5. It is important to let students express their ideas, even if they are wrong or absurd.

1 ☐ Strongly disagree | 2 ☐ | 3 ☐ | 4 ☐ Neither agree nor disagree | 5 ☐ | 6 ☐ | 7 ☐ Strongly agree

S6. It is more efficient to directly teach students the correct answers rather than asking them questions and spending time on their potentially incorrect responses.

1 ☐ Strongly disagree | 2 ☐ | 3 ☐ | 4 ☐ Neither agree nor disagree | 5 ☐ | 6 ☐ | 7 ☐ Strongly agree

S7. When a student asks a question about a topic that intrigues them, I only answer it if it relates to the subject I am covering at that moment. If it is not relevant, I postpone it so as not to disrupt the flow of the lesson.

1 ☐ Strongly disagree | 2 ☐ | 3 ☐ | 4 ☐ Neither agree nor disagree | 5 ☐ | 6 ☐ | 7 ☐ Strongly agree

D.6 Pedagogical Beliefs (Block 2)

Innate ability beliefs

S8. To succeed in school, working hard is not enough. One also needs a gift or innate talent.

1 ☐ Strongly disagree | 2 ☐ | 3 ☐ | 4 ☐ Neither agree nor disagree | 5 ☐ | 6 ☐ | 7 ☐ Strongly agree

S9. The most important factors for academic success are motivation and sustained effort; innate abilities are secondary.

1 ☐ Strongly disagree | 2 ☐ | 3 ☐ | 4 ☐ Neither agree nor disagree | 5 ☐ | 6 ☐ | 7 ☐ Strongly agree

Gender beliefs

S10. Although there are exceptions, boys are generally more gifted in mathematics than girls.

1 ☐ Strongly disagree | 2 ☐ | 3 ☐ | 4 ☐ Neither agree nor disagree | 5 ☐ | 6 ☐ | 7 ☐ Strongly agree

Teaching style

S11. A noisy classroom is not a problem as long as students are busy working.

1 ☐ Strongly disagree | 2 ☐ | 3 ☐ | 4 ☐ Neither agree nor disagree | 5 ☐ | 6 ☐ | 7 ☐ Strongly agree

S12. I do not like falling behind on the curriculum because of students' problems and questions.

1 ☐ Strongly disagree | 2 ☐ | 3 ☐ | 4 ☐ Neither agree nor disagree | 5 ☐ | 6 ☐ | 7 ☐ Strongly agree

S13. Students should be able to choose the activities we do in class.

1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐
Strongly disagree			Neither agree nor disagree			Strongly agree

D.7 Pedagogical Beliefs (Block 3)

Innate ability beliefs
S14. The most important factors for school success are motivation and sustained effort; inherent abilities are secondary.

1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐
Strongly disagree			Neither agree nor disagree			Strongly agree

Gender beliefs
S15. Although there are exceptions, boys are generally better at mathematics than girls.

1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐
Strongly disagree			Neither agree nor disagree			Strongly agree

Teaching style
S16. A noisy classroom is not a problem as long as students are busy working.

1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐
Strongly disagree			Neither agree nor disagree			Strongly agree

S17. To succeed in school, working hard is not enough. You must also have either an aptitude or talent of your own.

1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐
Strongly disagree			Neither agree nor disagree			Strongly agree

S18. I do not like falling behind the program because of difficulties and questions from students.

1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐
Strongly disagree			Neither agree nor disagree			Strongly agree