Discriminating Behavior: Evidence from teachers' grading bias

Ferman, Bruno and Fontes, Luiz Felipe

Sao Paulo School of Economics - FGV, Sao Paulo School of Economics - FGV

14 May 2020

Discriminating Behavior: Evidence from teachers' grading bias[*]

Bruno Ferman[†] Luiz Felipe Fontes[‡]

January 28, 2021

Abstract

Recent evidence has established that non-cognitive skills are key determinants of education and labor outcomes. However, little is known about the mechanisms producing these results. This paper tests a channel that could explain part of the association between some non-cognitive characteristics and educational attainment: teachers' assessment practices that unequally evaluate students on the basis of their classroom behavior rather than their scholastic competence. Evidence is drawn from unique data on middle- and high-school students in Brazilian private schools. Our main empirical strategy is based on the contrasting of teacher-assigned and blindly-assigned scores on achievement tests that are high-stakes and cover the same material. Using detailed data on student classroom behaviors and holding constant performance in exams graded blindly, evidence indicates that teachers inflate test scores of better-behaved students, and deduct points from worse-behaved ones. We also find that, conditional on end-of-year grade, teachers' decision to approve pupils that are bellow the passing cutoff grade is influenced by how these students behaved in class. Back of the envelope calculations suggest that this grading behavior may significantly change the proportion of students failing the school year depending on their classroom attitudes.

[*] The authors would like to thank Pedro Tepedino, Nicolás Ajzenman, and seminar participants at the European Winter Meeting of the Econometric Society, DSE Winter School, and FGV-EESP for helpful comments and suggestions. Luiz Felipe Fontes gratefully acknowledges financial support from Sao Paulo Research Foundation (FAPESP) Grant 2019/25473-5.

[†] FGV-EESP, bruno.ferman@fgv.br.

[‡] FGV-EESP, luiz.fontes@fgv.br.

1 Introduction

Researchers have emphasized that socially productive skills include not only traditionally studied cognitive abilities, but also behavioral and socio-emotional factors such as perseverance, self-control, and prosociality. In recent years, numerous studies have documented the central role played by these noncognitive skills in shaping educational attainment and adult outcomes (Segal, 2013; Heckman et al., 2006; Papageorge et al., 2019; Deming, 2017; Kautz and Zanoni, 2014; Heckman et al., 2014).[1] Importantly from a policy standpoint, there is also ample evidence suggesting that these skills are malleable, and can be influenced by school and teacher quality, home environment, and educational interventions (Jackson, 2018; Bertrand and Pan, 2013; Heckman et al., 2013; Jackson et al., 2020; Alan et al., 2019). However, despite the importance of socio-emotional factors, and significant advances in understanding its causes and consequences, there is still limited empirical evidence on *how* they affect important outcomes.

In the present study, we propose teachers' assessment practices as a potential mediator for gaps in attainment between students with different non-cognitive characteristics. We examine its prevalence by testing whether teachers unequally assess students on the basis of their classroom behavior rather than their scholastic competence.[2] The paper employs a unique administrative data from an educational company that manages more than one hundred private schools in Brazil. We use teachers' reports on their students' behavior to construct measures of good and bad in-class behaviors. Our main empirical strategy is based on the contrasting of teacher-assigned and blindly-assigned scores on achievement examinations that are high-stakes and designed to measure students' mastery of the same material. We also study teachers' discretion on whether to approve students with an end-of-year grade below the passing threshold and estimate if this is another high-stakes decision influenced by student classroom behavior. To do so, we compare students with the same final grade but with different behavioral characteristics.

We show that students' in-class behaviors substantially affect their scores in teacher-graded achievement examinations, even conditional on proficiency on the material required by the test. Good behavior produces a math test score bonus of 0.11 SD, and bad behavior a deduction of 0.14 SD. These estimates explain 20 and 40 percent of the unconditional correlation between good and bad behavior and test scores, respectively. These results are robust to a series of potential problems. To deal with the incidence of measurement error on the blind test scores used as regressors, we use lagged scores as an instrument for the current ones. We show that if the exogeneity condition of the instrument does not hold, our parameters of interest are

[1] For surveys, see Almlund et al. (2011), Heckman and Kautz (2012), and Heckman et al. (2019).

[2] We interpret classroom behaviors as narrower manifestations of personality characteristics. Several recent papers use observable behaviors from administrative data to measure non-cognitive skills and demonstrate this as a promising approach (Heckman et al., 2019). Some of them have used classroom behaviors based on teacher reports like we do (Segal, 2013; Papageorge et al., 2019; Heckman et al., 2013). Schooling behaviors have also been associated with traditional non-cognitive skills such as patience (Alan and Ertac, 2018; Castillo et al., 2011; Sutter et al., 2013), self-control (Duckworth and Seligman, 2005), and conscientiousness more generally (Segal, 2008). Moreover, Spengler et al. (2018) show that schooling behaviors have similar or even more significant predictive power of long-term outcomes than well-known psychometric measures of non-cognitive skills. Finally, Farrington et al. (2012) advocate that schooling behaviors are the main channel through which most traditional non-cognitive skills affect educational outcomes.

bounded by OLS and IV estimators under a few additional assumptions. A placebo test that uses as outcome blindly assigned scores on tests with the same format from the teacher-graded ones indicates that potential differences between blind and non-blind exams do not explain the pattern of our results. Also, exploiting blindly-assigned essay scores, we find no evidence supporting that our results are explained by potential biases from math teachers toward student handwriting skills. Our results are also robust to other confounders that may lead to grading bias, including gender and race. Moreover, we find much stronger grading bias due to in-class behavior than for these other features. Our results also remain similar if instead of using the behavior reports from the same teacher that assign math test scores, we use the behavior assessments from current teachers from other subjects or assessments made in the previous year, by teachers that currently do not even teach the students. Finally, teachers' grading behavior does not appear to be consistent with potential interpretations of statistical discrimination models: our results are more pronounced in classes where the correlation between ability and behavior is low; they are also stable throughout the year and across evaluations with different subjectivity levels. Taken together, our results suggest that students' behavioral characteristics directly affect teachers' assessment of their cognitive performance on achievement tests.

Consistent with the evidence on grading biases, we also find that, conditional on course grades, teachers decide whether or not to approve students below the official passing threshold based on how they behaved in class during the year. The well-behaved ones are in a higher proportion approved without having to take a reassessment examination. The opposite is true for the ill-behaved; in particular, those close to the passing threshold have a 20% higher chance of going through a reassessment phase instead of being approved directly by teachers. These results are robust to the omission of unobservable student characteristics that could also influence teachers' decision (Oster, 2019). Finally, using our main estimates to correct teachers' grading biases toward behavior, we find that the proportion of students who fail in the school year would be significantly different under counterfactual grading scenarios.

This paper provides quantitative evidence that teachers are not neutral to students' behavioral characteristics when assessing their current achievement and taking high-stakes decisions based on that. This grading behavior may be socially desirable if it induces a student to behave more positively, generating private benefits to the pupil and positive externalities to peers (Golsteyn et al., 2021). However, this practice is expressly prohibited by our partner's official guideline. This follows long-standing recommendations by specialists on classroom assessment, which explicitly warn against the adjustment of test scores to reflect students' attitudes and behaviors. (McMillan, 2013).[3] Moreover, there are two subjective grades in our setting that teachers may factor in their pupils' behavior. Hence, if teachers wanted to simply induce good behavior through grades, it would be more efficient to use these marks. These grades, however, have very low variability. Hence, consciously or not, teachers seem to use non-official channels to punish (reward) undesired (desired) behavior. This result is relevant to the understanding of educational gaps between students with different behavioral characteristics.

[3] 95% of the school administrators studied by Johnson et al. (2008) judged this practice as unethical.

Given the long-term consequences of grade retention (Manacorda, 2012; Eren et al., 2018) and inaccurate test scores (Terrier, 2016; Diamond and Persson, 2016; Lavy and Sand, 2018; Lavy and Megalokonomou, 2019; Dee et al., 2019; Nordin et al., 2019), our results may also suggest a potential mechanism for the relation between non-cognitive factors and other important life outcomes.

Our findings contribute to the literature highlighting the importance of non-cognitive skills. Little is known about the mechanisms behind the association between socio-emotional factors and educational and labor outcomes. Researches often speculate that non-cognitive characteristics induce productive habits which result in better life outcomes.[4] Evidence on that channel is mixed, depending on the skill being evaluated. While Lavecchia et al. (2016) show that impatient students report spending less time doing homework, other papers find that impatient students do not have lower study effort, even though they have much worse test scores (De Paola and Gioia, 2017; Non and Tempelaar, 2016). Outside the economics literature, a few papers show that personality traits are correlated with study habits (Lubbers et al., 2010; Credé and Kuncel, 2008). A related explanation is that non-cognitive characteristics may affect the effort put during tasks to obtain good results. Evidence by Borghans et al. (2008) support this explanation. The authors show that individuals with high non-cognitive skills operate a low-stakes cognitive test at a high level, even without rewards (see Segal (2012) for a related result). Similarly, Cubel et al. (2016) show that personality traits predict performance in an experimental task that requires real effort. The authors argue that this finding suggests that at least part of the effect of personality on labor market outcomes operates through productivity. Overall, the few evidence suggesting mediators for the association between non-cognitive skills and other outcomes are based on partial correlations and come from small-sample studies analyzing experimental outcomes. In our study, we analyze a significant number of students and make use of a quasi-experimental research design to provide evidence on a different mechanism – teachers' assessment practices – behind the relation between non-cognitive characteristics and schooling outcomes.

This paper is also closely related to the recent literature on teacher discrimination in grading. Some previous papers compare non-blindly- and blindly-assigned marks across minority and non-minority students (Botelho et al., 2015; Burgess and Greaves, 2013; Hanna and Linden, 2012; Alesina et al., 2018) and genders (Lavy, 2008; Hinnerich et al., 2011; Falch and Naper, 2013; Cornwell et al., 2013; Breda and Ly, 2015), and establish that grading discrimination exists in those dimensions. Besides ethnic and gender indicators, classroom behavior is another relevant student characteristic available to teachers during in-class interactions that may impact their judgment when grading. Mechtenberg (2009) refers to the behaviors as attitudes, which include

[4] Segal (2013) theorizes a similar channel. Her empirical results provide evidence that childhood misbehavior is negatively correlated with educational attainment and labor market outcomes. Based on the results of Castillo et al. (2011) – which find that pupils with higher discount rates have more behavioral problems in school – she develops a model to interpret the mechanisms driving her results. In her model, individuals are endowed with both cognitive and non-cognitive human capital. They can enhance their cognitive human capital at each level of schooling by exerting costly effort. Those who value the future less (i.e., those with low non-cognitive skills) invest less effort in school and hence accumulate less cognitive human capital; as a result, they earn less.

students' personality traits and habits that may change teachers' grading behavior depending on whether they like or dislike these attitudes. We contribute to the literature by testing this hypothesis in detail. Previous papers recognize that classroom behavior may be one of the most critical cofounders in grading discrimination estimates. Hence, a few of them try to adjust for proxies of behavior.[5] Similar to those papers, we also find grading discrimination toward ethnicity and gender. However, in this article, we use a unique dataset on objective indicators of student positive and negative in-class behaviors to show that, irrespective of ethnicity and gender, teachers score students' cognitive skills differently depending on how they behave in class. Moreover, we show that, besides the grading of examinations, there are other educational evaluation steps in which teachers make achievement assessments based on non-achievement characteristics.

Despite being one of the first studies to examine grading biases toward pupil behaviors in detail, we are well aware that the question of whether teachers factor the behavior of students into achievement assessments and grades is not new to the education and psychology literature. Previous research from education highlights that teacher assessments of student academic ability are central to school decisions, including instructional planning, screening, placement, referrals, and communication with parents. They can also directly influence study patterns, self-perceptions, and motivation to learn (Brookhart, 1997; Rodriguez, 2004). Even so, researchers working with classroom assessments have been warning about the unreliability of marks attached to pupils' work for a long time (Starch and Elliott, 1912, 1913). Many teachers misestimate students' abilities and misinterpret their performance in summative tests; frequently, they arrive at their assessment through idiosyncratic methods (Kilday et al., 2012). These tendencies may allow teachers to pull for students who deserve better grades or adjust scores down for students with poor behavior (Wyatt-Smith et al., 2010; Harlen, 2005). As far as we know, quantitative education researchers have not formally tested this hypothesis. Similar to the economics literature, they have focused on issues related to gender and race (Wen, 1979; Piché et al., 1977; Roen, 1992). This paper tries to fill this gap.

Research in psychology has studied the effects of socio-emotional and behavioral factors on test scores and grades extensively. Several studies from the seventies and eighties show that student temperament in the classroom strongly predicts teacher-assigned grades (Keogh, 1986). More recently, researches in psychology have evaluated the predictive power of student personality characteristics on grades and standardized achievement test scores (Almlund et al., 2011; Duckworth and Allred, 2012). An important finding from this literature is that, among the Big Five, conscientiousness – traditionally associated with student classroom behaviors

[5] Lavy (2008) adjusts for past grades under the hypothesis that those should be correlated with students' past behavior in the classroom, which should be correlated with the students' current behavior. Botelho et al. (2015) also use previous grades as well as several other variables. Among them, physical education grades, attendance records, and the perception of parents regarding their children's engagement at school. Cornwell et al. (2013) controls for "attitudes toward learning", which are based on teacher reports. Alesina et al. (2018) use a subjective grade decided jointly by all the teachers. Terrier (2016) uses a variable of disruptive behavior that equals one if the student received a disciplinary warning from the class council or if he/she was temporarily excluded from the school by the school head because of violent behavior.

(Segal, 2008) – is the most predictive skill of both course grades and test scores. However, a few papers show that some facets of conscientiousness seem to be better predictors of course grades than of achievement scores. Following our previous discussion, researchers argue that this may be the result of those abilities inducing more positive study habits, which translate into higher course grades. As achievement tests require the students to solve relatively novel problems, its scores may not reflect study habits as much as school-level works (Almlund et al., 2011).[6] Such a difference may also be the result of the influence of non-cognitive skills on the amount of effort that a test taker exerts in examinations with low stakes such as some standardized achievement tests (Heckman et al., 2019). Finally, it is also speculated that some socio-emotional factors may help students to behave positively in the classroom, which could be directly factored into course grades by teachers (Duckworth et al., 2012). In this article, instead of studying course grades and low-stakes standardized tests, we compare teacher-assigned and blindly-assigned scores from high-stakes achievement exams covering the same material. In our setting, study habits and effort should impact both types of evaluations similarly. Our study adds to this literature by providing empirical evidence that teachers rate students' competence in achievement tests differently whether they behaved positively or negatively in the class. As course grades are a direct function of these cognitive assessments, we suggest another reason why non-cognitive skills are important predictors for grades.

This article is organized as follows. Section 2 describes our data and presents our behavior measures. Section 3 presents our empirical strategy and main results on the grading biases toward student behavior. Section 4 presents robustness and heterogeneous analyses. Section 5 estimates whether approval decisions are based on student behavior. Section 6 simulates grade retention rates across students with different classroom behaviors under counterfactual grading scenarios. Section 7 concludes.

2 Data

2.1 Background and Data

We employ administrative data from a Brazilian private education company. The company manages more than one hundred private schools located in the South, Southeast, and Center-West of Brazil. Enrollment corresponded to more than eighty thousand primary, middle, and high-school pupils in 2018. To examine teacher assessment biases, we take advantage of administrative dataset that contains students' scores on tests graded blindly and non-blindly.

We now describe the schools' setting for students from middle school (grades 6-9) and from the first two years of high school (grades 10-11).[7] Schools operate a two-term school year (first and second semester). Each term is divided into three cycles. Students perform, per each

[6] Related to this channel, Borghans et al. (2016) find that IQ is a better predictor of SAT scores than of course grades.

[7] Appendix Figure A.1 complements this section by highlighting the main elements of our setting visually.

cycle, a “multidisciplinary exam” (ME) and a “specific exam” (SE). Both types of tests are high-stakes and do factor into the pupils’ end-of-term average score. Scores from SEs are worth 60%, while the scores from MEs are worth 30%. Students also earn a subjective grade based on their behavior on in- and extra-class activities, which factors 10% into the end-of-term average score.

MEs test knowledge on four topics: mathematics, language (English and Portuguese), science (physics, biology, and chemistry), and humanities (geography and history). All the questions from these exams are multiple-choice and corrected by a computer. The scores students obtain from these exams, which we call blind scores, can be assumed to be free of any bias caused by stereotypes from examiners. SEs are specific to each subject. Most of these exams are a mix of multiple-choice and questions requiring written answers, and are graded by teachers. By the schools’ rules, teachers’ grading should be only based on the students’ proficiency in the exam-specific subject. Still, teachers may have considerable arbitrariness when assigning grades. In that case, such scores, which we call non-blind scores, may be biased because of student behavior. In most of the schools, the third and the sixth SEs are only multiple-choice and also corrected by a computer. Additionally, in most of schools, students perform essays graded blindly by an external team. However, these essays are usually non high-stakes.

Both types of examination we described are created by the schools’ pedagogical team, based on a bank of questions. They are designed to measure students’ mastery of the material delivered within each cycle. Both tests are also taken under the same conditions: they take place in the students’ classroom and are supervised by inspectors, which are also responsible for giving general instructions. Depending on the school, the SE may be scheduled before or after the ME. The major difference between the tests is that the ME usually cover different subjects, while the SE exam is subject-specific. The exceptions are math and essay. As we aim to contrast blindly and non-blindly graded high-stakes examinations on the same subject, we will focus mostly on math scores. Still, we test for grading biases in other subjects as well.

To monitor students’ behavior, our partner developed a platform where teachers must report on a regular basis their pupils’ classroom behavior at some classes. Teachers must mark at least one of the following options when assessing their students’ behavior: “dedication”, “good interaction with classmates”, “participation during the class”, “excessive talking”, “cellphone use”, “disinterest during the class”, or “did not complete the required tasks”.

The dataset used in this study pertains to the school year 2019 and contains all the blind and non-blind scores of the students from middle school (grades 6-9) and the first two years of high-school (grades 10-11). We select these students as before grade 6, pupils do not perform blindly-graded exams. Also, in the last year of high-school (grade 12), students do not perform teacher-assigned scores. We also have access to the dataset coming from the schools’ online system, which contains all the behavior assessments teachers made in 2019 and 2018. We discuss next how we use these reports to construct behavior measures. Finally, our data also contain some major student characteristics: ethnicity, gender, and age.

2.2 Behavior Measures

In order to estimate both a potential discrimination against badly behaved students, as well as a potential favoritism toward the best behaved ones, we propose and compute two behavior measures. To do so, we start classifying the behavior reports into good and bad assessments. In particular, "dedication", "good interaction with classmates", and "participation during the class" are classified as an assessment of good behavior. Now, "excessive talking", "cellphone use", "disinterest during the class", and "did not complete the required tasks" are classified as an assessment of bad behavior.

Based on this classification, we construct measures of good and bad classroom behavior defined on the interval $[0, 1]$ for each student and each subject. The measure of good (bad) behavior weights the number of good (bad) assessments a student received from his/her teacher of a specific subject by the maximum number of good (bad) evaluations received by a classmate from that same teacher. These measures are formally defined as follows. Let $\mathcal{I}$ denote the set of all students. For any $i \in \mathcal{I}$, define $\mathcal{C}(i) \subset \mathcal{I}$ as the set of students in the same classroom of pupil i, including himself/herself. Let b_{ijs} and g_{ijs} denote the number of bad and good behavior reports received by i until exam j from a subject s teacher. The good and the bad behavior measures are defined, respectively, as:

$$GB_{ijs} := \frac{g_{ijs}}{\max\{g_{hjs} : h \in \mathcal{C}(i)\}},$$

and

$$BB_{ijs} := \frac{b_{ijs}}{\max\{b_{hjs} : h \in \mathcal{C}(i)\}}.$$

Notice that the good (bad) behavior measure from subject s is not defined for pupils in classrooms where $g_{hjs} = 0$ ($b_{hjs} = 0$) for all $h \in \mathcal{C}(i)$. These classrooms are discarded from our final sample. However, the GB (BB) measure is well defined for pupils with no good (bad) behavior assessments in a specific subject provided they belong to classrooms where at least one of their classmates received such an assessment. In that case, $GB_{ijs} = 0$ ($BB_{ijs} = 0$). These students can be understood as "neutral" with respect to the respective behavior measure. In robustness checks, we also test alternative ways of using the behavior reports: we use the behavior reports made by teachers from subjects other than s and by teacher from the previous year, currently not teaching the students.

2.3 Descriptive Statistics

Table 1 reports the summary statistics for our sample. The data cover 23,000 students from grades 6-11 in 738 classrooms and 80 schools. At least 58% of the sample is white, 24% is *Pardo*, 3% is black, and 1% is yellow or indigenous. A large share of parents (15%) did not provide their children's ethnicity. The gender split is roughly even. The average age is fourteen years old and 5% of the students are at least two years late in relation to the official age for their current grade.

In 98% and 99% of the classrooms, there are students with at least one bad and good behavior assessment, respectively. This share is lower if we consider only reports made by math teachers: 81% and 84%. Students received more positive than negative assessments. Indeed, they earned, on average, 44 assessments of good behavior and 12 of bad behavior. Of these, 10 and 3 came from math teachers. Moreover, considering all teachers' behavior reports, the students' average score in the good behavior measure is 0.53, while it is 0.29 in the bad behavior measure. The average score is similar if we only use the reports from math teachers. Appendix A presents additional statistics for the behavior data. Appendix Figure A.2 plots the empirical CDF of the behavior measures. The bad and the good behavior measures computed using the assessments from all teachers assume value zero for 17% and 1% of the students, respectively. The share of neutral students is higher if we consider only the assessments from math teachers – 35% and 10%, respectively. Appendix Figure A.3 plots the empirical CDFs by student demographics. Boys and over-aged students receive more negative behavior assessments and fewer positive behavior assessments. There is no clear association between ethnicity and in-class behaviors. Finally, Appendix Figure A.4 plots the association between the behavior measures and the subjective grades, which teachers may factor in their pupils' in- and extra-class behaviors. While it is clear that students getting the lowest subjective grades are also scoring high (low) in the BB (GB) measure, these grades have very low variability. Our goal in this papers is thus to evaluate whether teachers use non-official channels to reward or punish in-class behaviors.

Figure 1 displays the performance gap in teacher-graded achievement math tests between students with different behavior skills. Students at the top quartile of the GB measure distribution ($GB(Pct.75) = 1$) outperform those at the bottom across the entire achievement distribution (panel a). The opposite is true for students at the bottom quartile of the BB measure distribution (panel b). As teachers may have considerable arbitrariness when assigning grades, they may bias these cognitive assessments depending on whether they like or dislike the in-class behaviors of the students. To test whether this mechanism explains part of the achievement-gap depicted earlier, we will use blindly-assigned scores from exams covering the same material of the teacher-graded ones as the counterfactual test scores students would earn if there were no grading biases. Figure 2 plots the association between blind and non-blind math scores after we took their average across all the examinations. Both test scores are significantly correlated. One standard deviation (SD) increase in the blind scores is associated with increased teacher-assigned scores of 0.81 SD. Moreover, a linear model of the teacher-assigned test scores on the blindly-assigned ones explains 60% of the non-blind scores' variation. These numbers are expressive, especially if we consider test score measurement error. This is explored directly in the next sections.

3 Estimating grading biases toward student behavior

3.1 Empirical Strategy

We are interested in estimating a parameter of grading discrimination toward classroom behaviors, defined as the effect of in-class behaviors on test scores, conditional on the student proficiency in the subject and other characteristics that teachers may be biased at. To motivate our estimable equation, we assume a simple and intuitive statistical model for how test scores are defined. The non-blind scores of student i in exam j and subject s are determined by the following function:

$$S_{ijs}^{NB} = P_{ijs} + v_{ijs} + \Delta(W_{ijs}),$$

where P_{ijs} is student's proficiency. This component reflects factors such as i's knowledge in the subject s required by the exam j and his/her test-taking ability. The term v_{ijs} represents idiosyncratic factors, such as luck or how the student was feeling on a particular day, that are equal to zero in expectation. The term $\Delta(W_{ijs})$ represents potential bias by exam graders, who manipulate test scores based on student i's characteristics contained in W_{ijs}. In particular, we let

$$\Delta(W_{ijs}) = \beta' B_{ijs} + \phi' X_{ij} + \xi_{ijs},$$

where $B_{ijs} := (GB_{ijs}, BB_{ijs})$ is a vector of student i's classroom behavior in subject s computed using only behavior reports that preceded exam j – in some specifications it is computed using reports made by teachers from subjects other than s and by teacher from the previous year; X_{ij} includes ethnicity indicators (Black, Indigenous, *Pardo*, Yellow, and White), gender, age, and the past performance of student i in blind examinations;[8] and ξ_{ijs} include i's characteristics unobserved by the econometrician that teachers observe and may discriminate against.

We refer to P_{ijs} as the test score that i would receive in expectation if there was no grading bias. However, we only observe a noisy signal from it, coming from scores in examinations that cover the same content of S_{ijs}^{NB}, take place under very similar conditions, but are graded blindly (S_{ijs}^{B}), and, hence, are free of any kind of bias from the graders. We assume that

$$S_{ijs}^{B} = P_{ijs} + e_{ijs},$$

where the error term e_{ijs} may not be necessarily idiosyncratic. As S_{ijs}^{B} could potentially measure different skills, we can decompose

$$e_{ijs} = \tilde{P}_{ijs} - P_{ijs} + u_{ijs},$$

where $\tilde{P}_{ijs}$ is the i's proficiency required in the blindly-graded examinations and u_{ijs} is an idiosyncratic noisy. To make explicit that potential biases could arise if both types of examination

[8] In our main specifications we control for the cumulative average performance in the blind examinations from science, humanities, and languages. When not using IV, we also control for past performance in math.

were to measure different abilities, we linearly project P_{ijs} on $\tilde{P}_{ijs}$:

$$P_{ijs} = \delta \tilde{P}_{ijs} + r_{ijs},$$

so that r_{ijs} represents factors that are required only by the non-blindly graded examinations.

In a final step, as we pool the test scores from different examinations and classrooms, we add to our main specification classroom-by-subject fixed effects (α_{cs}) and exam fixed effects (π_j). Using previous definitions, we get that:

$$S^{NB}_{ijs} = \delta S^{B}_{ijs} + \beta' B_{is} + \phi' X_{ij} + \alpha_{cs} + \pi_j + \varepsilon_{ijs}, \tag{1}$$

where $\varepsilon_{ijs} \coloneqq \xi_{ijs} + r_{ijs} - \delta u_{ijs} + v_{ijs}$. Our parameter of interest is the vector β. It measures the effects of classroom behaviors on teacher-assigned test-scores, conditional on student's proficiency (proxied by S^{B}_{ijs}), other characteristics that teachers may be biased at (X_{ij}), and exploring only within classroom variation (α_{cs}). Its identification requires that we deal with unobserved heterogeneity ($\xi_{ijs} + r_{ijs}$) and measurement error in the blind scores (u_{ijs}).

We claim that, conditional on ethnicity, gender, age,[9] and past blind scores, if there other student characteristics available to teachers, varying systematically within the classroom, and affecting their grading behavior, they are balanced between students with different in-class behaviors. In particular, we are controlling for the characteristics the literature has shown teachers may be biased against. Moreover, we show our main point estimates are stable in specifications with and without controls for student characteristics, which may suggest that omitted variable bias is not a major concern if selection on observables is informative about selection on unobservables (Altonji et al., 2005; Oster, 2019).

We also assume that $\mathbb{C}(B_{ijs}, r_{ijs}) = 0$. We already discussed some particularities of our design that may provide grounds for the plausibility of the assumption. Overall, we believe there are no apparent systematic differences in the exam-taking environment that could interact with i's characteristics. Both the blindly and the non-blindly graded exams are school-level tests that take place in the regular classes and are supervised by inspectors. Furthermore, both types of exams are high-stakes and designed by the pedagogical sector to cover the same material, based on a bank of questions. Our main concern is that, while the MEs are only multiple-choice, most of the teacher-graded exams are a mix of multiple-choice and written questions. If subjective questions require handwriting abilities not covered by the blindly-graded exams, correlated with in-class behaviors, our identification strategy could not be valid. To investigate whether this seems to be a potential concern, we make use of a reduced sample of schools where students perform essays graded by an external team. Xu and Gong (2017) show that blindly-assigned essay scores reflect grading biases toward handwriting quality. Therefore, we control for blind essay scores in an attempt to control for students' writing skills and potential grading biases toward handwriting abilities. We find nearly identical results when adopting such strategy.

[9] As we exploit within classroom variation, age should capture students who are advanced or who have failed previous years.

Additionally, we show that essay teachers practice grading discrimination toward classroom behavior. In this case, we are able to present evidence of grading biases in a setting where both the blind and the non-blind exams have exactly the same format. Finally, to test if there is any other difference between SEs and MEs correlated with behavior that could bias our estimates, we additionally use the scores from the SEs that are blindly graded as placebo outcomes. In particular, we show that such scores are not affected by student behaviors conditional on the blind scores.

To tackle the measurement error problem, we use the lagged blind math score (LS_{ijs}^{B}) as an instrument for the current one. In the literature on grading discrimination, this is the same strategy of Botelho et al. (2015). In the context of value-added models, simulations by Lockwood and McCaffrey (2014) suggest that using lagged scores as instruments for the current ones can eliminate the bias in treatment effects estimations when test scores used as regressores are measured with error.[10] Still, one might be worried about the validity of the exclusion restriction. It might be, for example, that teachers practice statistical discrimination by using students' past performance in blind math exams to reduce noise about their proficiency. The exclusion restriction may be valid provided we adjust especially for past test scores in other subjects, and also for pupil's ethnicity, gender, age, classroom behavior, and classroom fixed effects. Otherwise, it is likely that lagged blind scores would be correlated positively with the unobserved skills that determine S_{ijs}^{NB}. Under this scenario, we show in Appendix B that OLS and IV produce upper and lower bounds for β. Due to test scores measurement error, the bias of the OLS estimator of β is bounded away from zero. The intuition is that behaviors measure part of δ through the correlation between behaviors and S_{ijs}^{B} once δ is estimated with attenuation bias. Contrary, if $\mathbb{C}(LS_{ijs}^{B}, \varepsilon_{ijs}) > 0$, δ is overestimated by IV, and hence, β is estimated with attenuation bias.[11] Anyway, when we estimate β by OLS and consider additional proxies for P_{ijs}, we obtain upper bounds that are close to the IV estimates. Additionally, we obtain very similar results if we follow other papers from the literature and estimate a differences-in-differences specification, which is equivalent to imposing the restriction $\delta = 1$ in equation (1) so that we do not need to deal with the measurement error problem.

Regarding inference, standard errors are robust to heteroskedasticity and calculated with student-level clusters. We also tested for school-level clusters and the standard errors remained nearly identical.[12] Additionally, in all our specifications, the test scores are standardized to a

[10] The authors propose several alternative methods based on the standard error of the test score – returned by Item Response Theory (IRT) – to correct for the measurement error bias. As the school examinations we study are not constructed using IRT, we can not use these methods. However, other papers that also use lagged scores as instrument to correct for the measurement error report that doing so using the standard errors returned by IRT leads to very similar results (Khwaja et al., 2011; Botelho et al., 2015). Another important finding by the authors is that it is possible to mitigate the influence of test measurement error in OLS regressions by controlling for multiple prior test scores. We also test such a specification by including in X_{ij} past cumulative performance in other subjects: language, science, and humanities.

[11] $\mathbb{C}(LS_{ijs}^{B}, \varepsilon_{ijs}) > 0$ if the lagged blind scores are correlated positively with the unobserved skills that determines S_{ijs}^{B} – $\mathbb{C}(LS_{ijs}^{NB}, \xi_{ijs} + r_{ijs}) > 0$ – and the measurement errors are not auto-correlated or the serial correlation is lower in comparison to $\mathbb{C}(LS_{ijs}^{NB}, \xi_{ijs} + r_{ijs})$.

[12] Appendix Table B1 presents this result.

distribution with zero mean and a unit standard deviation. This procedure is applied within subjects to each test separately. To facilitate reading of results, in our main specifications B_{ijs} stands for binary variables that indicate whether students are in the top quartile of the behavior measures' distribution. We also present the results using the continuous measures. To avoid feedback effects, our behavior measures use only reports that preceded the teacher-graded examinations. We also tested the reports students received in the previous year by teachers that currently do not teach them.

3.2 Main Results

We begin by examining the association between behaviors and test scores. Table 2, column (1), presents the unconditional OLS estimates. We can see that the average math grades of students with bad in-class behaviors – $BB(Pct.75) = 1$ – are 0.42 standard deviation (SD) below those such that $BB(Pct.75) = 0$. The unconditional grade gap between students with $GB(Pct.75) = 1$ and $GB(Pct.75) = 0$ is even greater: 0.65 SD in favor of the better-behaved pupils. As several other non-cognitive indicators studied by the literature, our behavior measures strongly predict pupil test scores. Of course, this does not imply directly that teachers are practicing grading discrimination. If, for example, students with better in-class behavior are those who prepare more for the tests, we should expect them to obtain higher grades.

Therefore, in column (2), we follow the empirical strategy previously outlined. We control for the blind math scores, our proxy for the grades students would obtain if there were no grading bias. To tackle the measurement error problem in our control variable we instrument current blind scores with its lagged values.[13] Under this specification, the behavior effects are significantly reduced, indicating that a share of the competence differences seen by teachers is captured by performance in the blindly-scored tests. Still, the behavior effects are significant and high in magnitude, indicating that teachers discriminate student behaviors in grading. Our results suggest that the better-behaved students have their grades inflated by 0.12 SD. This amounts to 18 percent of the unconditional gap. Additionally, teachers seem to deduct, on average, 0.16 SD from worse behaved students, which represents 38 percent of the unconditional gap. Just to put it into perspective, a recent meta-analysis of field experiments in education found that 70 percent of the treatment effects on math test scores produced by 314 RCTs are lower than 0.16 SD (Kraft, 2020). Hence, grading biases toward classroom behavior may produce similar effects from successful interventions in the educational domain. Additionally, these biases produce a gap between students with different behavioral skills that, in our context, amount to approximately 60 percent of the black-white achievement gap (0.23 SD).

Our results remain similar if we control for student demographics (column 3). Hence, despite the vast literature showing grading biases toward boys and minority ethnic groups, they are not relevant cofounders in our setting. In the Appendix C, we present evidence suggesting

[13] Reflecting the cumulative nature of student performance, past scores are strongly correlated with current ones as it is suggested by the high first-stage F statistic. Additional first-stage summary statistics are available upon request.

discrimination against boys and black students, in line with evidence from the literature. Results are, though, much weaker than the results we found for behavior. Additionally, we do not find heterogenous effects according to these characteristics, making it unlikely that our results are explained by teachers' biases toward gender and ethnicity also reflected on the behavioral report. Grade repetition is another characteristic that might influence teachers' grading decisions. As we explore only within classroom comparisons and our results are robust to the inclusion of a set of age indicators, this does not seem to be a relevant cofounder either. Results are also robust if instead of controlling for a set of age indicators, we add a covariate capturing age-grade distortions. In column (4), we control for past blind scores in other subjects, which may proxy for unobservable competences required only by the non-blindly graded exams. Finally, in column (5), we follow the same specification of Botelho et al. (2015), which also correct for the measurement error in language test scores, and consider higher-order polynomials for the blind scores.[14] In particular, a cubic polynomial for the blind math scores, a linear function of the blind language scores, and the interaction between these. Under these specifications, the bad behavior estimate reduces slightly, although the reduction is not statistically different from zero, and the good behavior estimate remains the same.

Appendix Figure A.5 replicates our previous results for several other subjects. This is consistent with other papers investigating teacher grading biases (Lavy, 2008; Hanna and Linden, 2012; Lavy et al., 2018). Additionally, previous evidence does not depend on our discretization of the behavior measure. When we use other moments of the distribution and the continuous measures, results reveal the same pattern (see Appendix D.1 for further results). We also analyze the different behaviors separately. Disinterest during class seems to be the negative behavior that influence teachers' grading the most. Participation during the class seems to be the most praised one (see Appendix D.2). Finally, we also use alternative data from a small number of schools where students take regular courses aimed at improving their socio-emotional abilities (more specifically, grit, creativity, cooperation, communication, pro-activity, and critical thinking). The pattern of the results remain similar if we proxy students' behavior skills using the scores they earn in these courses: 35% of the unconditional correlation between non-cognitive skills and scores on teacher-graded achievement tests seems to be mediated by grading biases toward socio-emotional characteristics (for more details, see Appendix D.3)

Taken together, our results suggest that teachers unequally grade students' academic cognitive skills on the basis of their classroom behavior rather than performance. The estimated grading biases are high in magnitude and affect students' scores similarly to successful educational interventions. Our results indicate that teachers' discretion in grading may explain a significant share of the positive association between non-cognitive characteristics and academic performance, systematically reported by the literature.

[14] Instrumenting language scores may be especially important if one believes that language skills are required only by the non-blind math scores.

4 Robustness and heterogeneity

In this section, we conduct additional robustness checks and empirically assess alternative explanations for our results.

4.1 Alternative estimators

As previously discussed, if the exclusion restriction of our IV estimator does not hold, then we should expect that we underestimate the behavior effects. We then estimate equation (1) by OLS to obtain upper bounds for the true discrimination parameter of interest β. Table 3 presents the results. In column (1), we replicate the same specification from Table 2, column (2). The first thing to notice is that the relation between blind and non-blind scores estimated by OLS (0.42) is much lower than the estimated by IV, reflecting the attenuation bias due to test score measurement error. As a consequence, the behavior effects are higher when estimated by OLS. Another major difference is that the inclusion of past scores in column (4) changes significantly the magnitude of our estimates. The reason is that the past blind scores from other subjects serve as proxies for part of the student proficiency signal, which mitigates the biases from the estimated behavior effects (Lockwood and McCaffrey, 2014). This allows us to estimate finer upper bounds for our parameter of interest. We also tried to reduce the biases even more by using past math scores as control, instead of instrument, but the results remained virtually the same (column 5). Overall, we highlight that the results estimated by OLS using additional proxies for student proficiency (-0.17 and 0.20 for the BB and GB measures, respectively) are close to the IV estimates (-0.14 and 0.12, respectively), and hence underestimation of β does not seem to be a concern. Finally, column (6) also estimates equation (1) by OLS but restricts δ to 1. In this case, our dependent variable becomes the difference between blind and non-blind scores and we do not have to deal with test score measurement error. Results are nearly identical those we obtain using IV.

4.2 Differences between blindly graded and non-blindly graded examinations

Our main empirical strategy relies fundamentally on the proximity between teacher- and blindly-graded achievement tests, which we claim to be granted in our setting. Under the model outlined in Section 3, δ represents the association between the skills required by both types of exam. The estimated δ is not statistically different from 1 when we correct test score measurement error using the IV strategy (Table 2, column 4). Such evidence gives empirical support to our claim. Next, we exploit particular features of our setting to conduct empirical tests and heterogeneous analysis that provide additional support to our study design's internal validity.

4.2.1 Placebo test

We argued previously in the text that there is no systematic difference between MEs and SEs in the exam-taking environment that could interact with behavior and performance. Furthermore, that the few apparent differences between them should not change the way we interpret the results. Specifically, we believe that there should not be relevant cofounders associated with the fact that SEs have a relatively higher weight in the course grades and that both exams are not taken simultaneously. To provide evidence in favor of our assumptions, we conduct a placebo test exploiting that the third and sixth SEs are blindly graded in most of the schools. More precisely, we estimate the effects of positive and negative behaviors on the blindly-assigned SEs' scores, conditional on the respective MEs' scores. In this case, as both test scores come from blindly graded examinations covering the same material, we do not expect behavior to have any effect unless there are relevant differences between the two kinds of exams. Table 4 presents the results. Supporting our claim, the correlations between bad (-0.2 SD) and good behavior (0.5 SD) and math test scores (column 1) vanish when we adjust for student proficiency measured by blind scores (column 2).

4.2.2 Subjective questions

We believe the only cofounder not ruled out in the previous test is handwriting ability. One might suspect that math teachers praise good handwriting when grading questions that require written answers. If these skills correlate with behaviors, our estimates would be biased. To test this hypothesis, we re-estimate our main results using a subsample of schools where students perform blindly-graded essays, and then adjust for the blind essay scores to check whether this seems to be an important confounder. This kind of examination, more than any other, should capture abilities like those mentioned before. Appendix Table B2 presents the results. In summary, we find nearly identical grading biases when we control for the essay scores, even if we correct its measurement error by using lagged scores. We also use the blind essay scores to test whether teachers from this subject practice grading discrimination toward student behaviors. In this particular case, we were able to compare evaluations with exactly the same format. The results we obtain for essay are very similar to those obtained when analyzing math scores (see Appendix Table B3). Thus, we are confident that our main results are not biased due to writing competencies praised by teachers in math achievement tests. This is further evidence suggesting that the few differences between MEs and SEs do not explain the pattern of our results.

4.2.3 Timing of exams

The different timing of the MEs and SEs could, in theory, account for some of the gaps in performance between students with different behaviors in the blindly and non-blindly graded exams. For example, one could argue that while students with worse classroom behavior tend to rest between the first and the second exam, those better behaved study. In this case, performance in the blindly graded tests may not reflect precisely the students' level of proficiency at the time they take the non-blindly graded tests, especially if the time gap is large. As for most of the schools, the teacher-graded examinations take place after the MEs (Appendix Figure A.6), this could explain the pattern of our results.[15] Previous evidence on placebo outcomes already suggest that different timing should not be a driver of our results. Still, as we have timing heterogeneity in our data, we provide a further piece of evidence supporting this result. Appendix Table B4 presents the results for two samples that differ on whether teacher-graded examination take place before or after the blindly-graded ones. Appendix Table B5 exploits variation in the intensive margin: in one sample the difference between the two types of exam is less than 3 weeks; in the another one, the opposite is true. The estimated biases have the same sign and are high in magnitude in all samples. These results suggest that the specific pattern in the timing of the blind and non-blind exams is not the cause of the pattern in our main results.

4.2.4 Student ability

Our results may have a different interpretation if there are still facets of student cognitive attributes affecting blind and non-blind scores differently, not ruled out by our previous robustness tests. To shed light on this possibility, we follow Lavy (2008) and control for student unobserved ability using academic performance in the previous school year. More specifically, we use the average score in all blindly-graded tests made in 2018. Appendix Table B6 reports such results. Our estimates remain remarkably similar when we include controls for math and language ability. This evidence reinforces that our results are not driven by unobserved competencies correlated with behavior required by only one type of examination.

4.3 Biases in the behavior reports

A potential concern in our setting is that behavior reports and test scores may have correlated attribution errors. When sending a behavior report, teachers may miss-attribute it due to other students' characteristics and attitudes besides classroom behavior. If they also use the same factors to discriminate test scores, our identification strategy would not be valid. We already showed that our results are not heterogeneous across sub-samples that vary according to gender and ethnicity, characteristics the literature has shown teachers may be biased against. Still, there may be unobservables that teachers project on both test scores and behavior reports.

We then measure student classroom behavior by using the reports from teachers other than the ones assigning math test scores. We first use the assessments made by all teachers ("All Teachers"). We believe these depend less on the subjectivity of a teacher's type, hence capturing better student behaviors and not other unobservables that could also be factored into test scores. We also test the sensibility of the results when we exclude the reports sent by the

[15] One could imagine several other mechanisms. Some, however, would not explain our results. For example, if students with worse classroom behavior tend to study for the exam later than well-behaved pupils, perhaps because they have a higher discount rate as suggested by the literature, than they might be better prepared for the second exam than for the first.

math teachers ("All Teachers – Math").[16] Finally, we exploit the fact that students do not have classes with the same teacher for two straight years in our schools. As we have access to the 2018 behavioral data, we can then measure the students' in-class behavior using assessments made in the previous year, by teachers that currently do not even teach the students. Figure 3 presents all the results. They remain remarkably similar across the behavior measures. Indeed, Appendix Table B7 shows none of the five different ways of estimating behavior leads to differences that significantly weaken our results. This may alleviate concerns with correlated attribution biases on behavior reports and test scores. Still, we cannot rule out that several different teachers may project the same students' unobservables, available to teachers in both 2018 and 2019, in the behavior reports and test scores.

4.4 Statistical Discrimination

Here we exploit data heterogeneity to test whether teachers' grading behavior is consistent with potential interpretations of statistical discrimination models.

4.4.1 Biases when ill-behaved students are skilled in math

Statistical discrimination arises if teachers use classroom behavior to predict students' unobserved ability based on beliefs they have on the proficiency of students with certain characteristics. The question of whether such beliefs are based on real evidence or are unfounded is irrelevant to the outcome of statistical discrimination. However, it is plausible to imagine that these beliefs may be influenced by the superior average performance of students with better classroom behavior.

We explore heterogeneity in the performance of students in the blindly-graded examinations to test whether the estimated biases are lower in a subsample of students where those with worse classroom behaviors are as skilled as their classmates with better behaviors. To do so, we select classrooms where in the first semester students with $BB(pct.75) = 1$ or $GB(pct.75) = 0$ performed, on average, better in the blind math exams than their classmates with $BB(pct.75) = 0$ or $GB(pct.75) = 1$, respectively. We call this subsample of sample A. The subsample where the previous conditions are not satisfied is called of sample B. Appendix Figure A.7 shows that in sample A, In sample B, there is a striking gap between these groups of students.

We believe those gaps should influence teachers' information about students' scholastic ability. Directly, if teachers do observe the students' performance in blindly-graded examinations, or indirectly, if proficiency in mathematics captured by performance on those examinations is correlated with performance in other outputs available to teachers during the year. Figure 4 presents the estimated biases for the full sample, sample A, and sample B. The estimates are though higher within sample A, which is the opposite of what one should expect under statistical discrimination based on unbiased beliefs. These results may suggest that this form of discrimination is not explaining our results. Otherwise, teachers are not updating their beliefs

[16] We also tested the reports coming from all the other subjects separately. The results are very similar across different subjects.

based on the average performance of the math class they are grading.

4.4.2 Subjectivity in evaluation

According to Bohren et al. (2019), the level of subjectivity in evaluations provides some piece of evidence that may help to disentangle sources of discrimination. More specifically, decreasing the subjectivity of evaluations should mitigate statistical discrimination, as beliefs about group statistics (e.g., the average scholastic ability across groups of students with different behaviors) play a smaller role in assessing quality when signals of quality are more precise (i.e., there is less subjectivity). Subjectivity should not affect preference-based discrimination, though, which will persist even if quality is perfectly observable. In this paper, we found that classroom behavior influence teachers' ratings of cognitive skills in several subjects, ranging from math and science to Portuguese and English (Appendix Figure A.5), and even essay (Appendix Table B3). We expect the level of subjectivity to be different across these evaluations. As our results have low heterogeneity within this dimension, teachers grading behavior in our data does not seem to be consistent with statistical discrimination.

4.4.3 Learning across the school-year

We now explore dynamic predictions of statistical discrimination models. Bohren et al. (2019) show that in settings where individuals repeatedly perform tasks that generate output, observing sequences of evaluations from such tasks should mitigate statistical discrimination, which can even be reversed if stemmed from inaccurate beliefs. Similarly, in the employer-learning model (Altonji and Pierret, 2001), employers observe workers' performance on the job and thus learn about workers' unobserved productivity. As they learn, they rely less on easily observed workers' characteristics to predict their productivity. The faster employers learn, the shorter the period during which firms statistically discriminate. In our setting, under statistical discrimination, the more teachers observe their students' performance at school, the less they should use classroom behaviors to predict students' proficiency.[17]

Exploring the fact that we observe teachers grading at least four examinations, we test whether estimated biases decay throughout the year. When lagged blind scores are used as instrument for the current scores, we lose information from the first exam. Additionally, the subset of schools in which the third and sixth SEs are teacher-graded is too small, so we cannot precisely estimate exam-specific grading biases for these cases. Then, Figure 5 presents the results for the second, fourth, and fifth exams. The estimated grading bonus for well-behaved students are remarkably constant across the year. The grading deduction from the ill-behaved ones is lower in exam 2 and then remains similar across exams 4 and 5, which may indicate that teachers take some time to identify students with bad behavior. Anyway, the results are the opposite of what one would expect under our initial hypothesis. These evidence suggest that

[17] In the context of racial grading discrimination, Botelho et al. (2015) find that while there is a bias toward black students attending classes with a teacher for the first time, no significant disparities are found among those that have already had classroom with that teacher before.

a model of statistical discrimination with learning does not seem to be explaining our results, perhaps because teachers do not statistically discriminate, or because learning is slow in our setting.

5 Estimating whether approval decisions are based on student behavior

So far, we have documented that students' in-class behaviors affect their scores in teacher-graded achievement examinations through a channel other than the material's knowledge required in those tests. The estimated effect is high in magnitude and is not explained by test score measurement error, differences between teacher-graded and blindly-graded exams, and a potential simultaneity between non-blind scores and behavioral reports. These results suggest that teachers may be influenced by how their pupils behave in class when scoring them in achievement tests, which is explicitly prohibited by the schools' rules. As shown in the previous section, this grading behavior is systematic throughout the school year. Hence, it may significantly affect students' academic progress. In this section, we try to go one step further and analyze if teachers also impose different standards for students in the final stage of the schooling evaluation process, depending on how they behave in class. More specifically, we study whether teachers may have discretion over the decision to approve students at the end of the school year and whether such a decision prejudices/benefits students depending on their classroom behaviors.

By our partner's official guidelines, students who achieve an end-of-year grade above 60 move on to the next grade. Those that do not meet the approval cutoff should be reassessed. This process is based on a reassessment examination that covers the entire content of the year, worth 100 points, and takes place approximately two weeks after the students' last examination. A student in the reassessment phase is approved if the simple average of her final grade and reassessment grade exceeds 50. However, in our data, several students with an end-of-year grade below 60 receive the privilege of being approved without going through the reassessment phase (see Appendix Figure A.8). After the last examinations of the second semester, teachers meet to discuss their pupils' performance. In most schools, teachers use these meetings to approve, with no need for reassessment, students with a final grade between 40 and 60.[18] There is no official rule that determines which students can pass under these conditions. Anecdotal evidence suggests that classroom behavior is a frequent argument used by teachers to justify their choices. We wish to estimate whether this behavior does exist and is systematic.

By plotting the proportion of individuals taking the reassessment exam across end-of-year grades and by in-class behavior status, Appendix Figure A.9 indeed suggests that ill-behaved students are harmed while the well-behaved are benefited. To quantify this difference and exploit

[18] Those meetings are called *conselho de classe* (literally translated by class council) and they are very common in Brazilian school setting.

cleaner variation – conditional on other students' characteristics and within the same grade and school – we propose the following empirical strategy. We pool i's end-of-year grades from each subject s and estimate for each final grade bin of size 1 between 41 and 59 the following specification:

$$R_{is} = \alpha' B_i + \gamma' W_i + \eta_l + \lambda_g + \psi_s + \epsilon_{is} \tag{2}$$

where R_{is} equals one if i took reassessment at subject s and zero otherwise (there is, if he was approved without having to pass thorough the reassessment phase). $B_i := (GB_i, BB_i)$ based on all reports student i received. As the decision to place low-grade students into the reassessment phase is taken together by teachers of different subjects, our behavioral measures use the reports from all of them. Still, the results remain similar if we use subject-specific measures. Results are also robust if we use the behavior assessments from 2018, made by teachers currently not responsible for approval decisions. W_i is a vector including age-grade distortion, gender, and ethnicity indicators. Finally, η_l, λ_g, and ψ_s are fixed effects for school, grade, and subject, respectively. Our vector of interest is α, which measure the correlation between in-class behaviors and the outcome of interest R_{is} among students with the same end-of-year grade, conditional on a few demographics, and exploiting only within school and grade comparisons. Regarding inference, standard errors are robust and clustered at the student level. As the estimation of α exploits variation within very restricted cells, one might worry about over-rejection of the null due to a potential low number of clusters within each behavior status. Following Ferman (2019), we find that the assessment for a 5%-level test is less then 6% for each grade bin, for both good and bad behavior indicators.[19] This indicates that inference based on clustered-robust standard errors at the student level is reliable.

Figure 6 plots our main estimates. Consistent with the pattern of results depicted in the previous sections, well-behaved students are systematically favored, while the bad-behaved ones are punished. The former are in a higher proportion approved without having to take one final examination; the opposite is true for the latter. The effects are stronger for students with a final grade closer to 60 (between 53 and 58). In this case, the point estimates indicate an increased or decreased chance of going through reassessment of almost 10 p.p – 20% in comparison to the rate of students going through reassessment within this range. Points estimates have a lower magnitude at the bottom of the grades' distribution as the outcome has very low variability in this case. A summary parameter computed by the average (weighted by bin size) of the effects across the grade bins indicates an effect of approximately 5 p.p for both groups. This represents 7% of the proportion of students taking reassessment within this sample. We do not find such an effect for other student demographics (see Appendix Figure A.10). Overall, these results indicate that teachers pose another obstacle to students' progress because of their classroom behaviors.

Even if our strategy fully control for the main determinant of reassessment and we further adjust for student demographics and fixed effects, there may still be some unobservable heterogeneity within each cell. In that case, the correlation we estimate may also measure the

[19] We consider iid standard normal variables in this assessment. Results are available under request.

relation between R_{is} and unobservable characteristics (associated with in-class behavior) that teachers may consider when deciding whether to approve low-grade students. We then follow Oster (2019) and use the change of our estimates in specifications with and without controls (including the fixed effects) scaled up by changes in the r-squared to see whether our results are robust to unobserved heterogeneity.[20] Instead of testing the stability of each grade-bin estimate, we test the robustness of the average effect. Appendix Figure A.11 shows that there is a slight reduction in the average effect for ill-behaved students when we adjust for student demographics and fixed effects. The average effect of good behavior remains stable. Following Oster (2019) (and her notation) we then assume that the r-squared of the model with covariates ($\tilde{R}$) would increase 30% if we could further include the unobservable determinants of R_{is} in the model ($R_{max} = 92\%$ in our setting) and that the unobservable and observables are equally related to the treatment ($\delta = 1$). Using this parametrization, we estimate the bias-adjusted treatment effect, depicted in Appendix Figure A.11 under the label "unobservables". Overall, the estimates would remain stable if we could control for unobserved heterogeneity. If we consider $\delta = 1$ as an upper bound scenario, the average effect of bad behavior on reassessment within the sample that may receive the privilege of being approved without the final exam is within the interval $[0.032, 0.045]$. The good behavior average effect would be within the interval $[-0.04, -0.036]$. These intervals would include the zero only if $\delta = 4$ and $\delta = 10$, respectively, which are very unlikely situations.

6 Simulating how assessment biases affect grade retention

In the previous sections, we documented that there exist different layers of discrimination against student behaviors. Test scores from several teacher-graded achievement examinations overweight or underweight students' knowledge in the subject depending on how they behaved in class. They are also treated differently in approval decisions at the end of the year. We now ask whether this teacher grading behavior affects grade retention probability depending on the students' behavior. To do so, we use our previous estimates and simulate counterfactual scenarios where this grading behavior would affect retention mechanically by altering students' end-of-year grades and their probability of going through the reassessment phase. Though, we have in mind that this would be only a conservative scenario, as discriminatory grading behavior may affect students' outcomes through several other channels not measured here as distorted self-perceptions (like reinforced beliefs of inferiority) and distorted perceived returns of human capital investments by parents.

We begin by noting that in our setting the probability of being retained is given by

$$\Pr(Retention = 1) = \Pr(Retention = 1 | Reassessment = 1) \times \Pr(Reassessment = 1).$$

[20] To do so, instead of estimating a regression for each grade bin, we estimate a satured model of R_{is} on grade fixed effects ant its interactions with the other independent variables of model (2).

Based on our main estimates from Section 3, we can correct or intensify teachers' grading biases toward classroom behavior to analyze how this affect $Pr(Retention = 1)$ through distorted final grades. Notice that end-of-year grade affects both $\Pr(Retention = 1|Reassessment = 1)$ and $\Pr(Reassessment = 1)$ as it is the main determinant of reassessment, and it is also factored into the scores that determine retention. Discriminatory attitudes toward classroom behavior may also affect $\Pr(Retention = 1|Reassessment = 1)$ through other channels, such as grading discrimination in the reassessment exams. However, as we do not have a clear design to estimate such a bias, we consider a lower bound scenario in which the only channel by which teachers' biases toward behavior affect this term is through its effect on final grades. Our current setting is a bit unusual as there are several high-stakes examinations graded blindly, which minimizes the effects of teachers' discretion when grading exams on students' end-of-year grades. Therefore, we simulate a more realistic baseline scenario where all test scores would be affected by grading biases toward behavior ("non-blind setting"). Then, to analyze how grading discrimination affects retention and reassessment, we simulate a scenario with no grading biases toward classroom behavior ("blind setting"). In this case, we also remove leniency to approve low-grade students based on classroom behavior, which directly affects $\Pr(Reassessment = 1)$.

Figure 7 plots our simulations on the probability of going through the reassessment phase. Among the well-behaved students (panel a), the proportion taking the reassessment exam is 8.6%. Everything else constant, the proportion of well-behaved students in the reassessment phase would be 1 p.p. lower if all tests were subject to grading manipulation. From here, if we could eliminate grading discrimination and also teachers' discretion on promotion decisions, the proportion taking reassessment would increase 3.6 p.p., or 40% in relation to the fraction in the current setting. The proportion of ill-behaved students (panel b) doing the reassessment exam is much higher: 36.7%. *Ceteris paribus*, it would be 40.6% if all test scores were affected by grading discrimination. We calculate that this fraction would decrease by one quarter (to 30%) if there were no grading biases toward in-class behavior. Our simulations suggest that the level of requirement imposed on students strongly depends on how they behave in the classroom. While the difference in the likelihood of going through reassessment between good- and ill-behaved students would be 33 p.p in the non-blind setting, it would be almost three fifths of this value (19 p.p.) in the blind setting.

Figure 8 then depicts our simulations for the proportion of retained students. Comparing the scenario in which all test scores were subject to grading bias to a scenario in which there is no grading bias, we find that the proportion of well-behaved students falling the grade would increase more than twice, from 0.17% to approximately 0.48%. Among the ill-behaved, the retention probability would go from 6.4% to 3.6% if teachers were neutral to students' classroom behavior when evaluating their achievement. We conclude from these simulations that teachers' discretion may impose or reduce significant barriers to students' academic progress contingent on their classroom behaviors. Teachers' assessment practices may then explain part of the relation between those non-cognitive characteristics and schooling variables, and even other important life outcomes directly affected by grade retention (Manacorda, 2012; Eren et al., 2018).

7 Conclusion

In this article, we empirically detect that teachers are not neutral to students' behaviors when assessing their aptitude and making high-stakes decisions based on that. When rating students' performance on achievement tests, teachers overweight or underweight their scholastic cognitive skills depending on how they behave in class. Approximately 20% (40%) of the unconditional correlation between negative (positive) classroom behaviors and teacher-assigned scores seems to be mediated by grading biases. These results are robust to the incidence of measurement error on blindly-assigned scores used as regressors, differences between blindly- and non-blindly- graded exams, student unobserved ability, and potential feedback effects between teacher-awarded scores and behavior assessments. Heterogenous effects suggest that grading biases are unlikely to result from statistical discrimination: they do not decrease in classes where the correlation between behavior and ability is low; they are similar across evaluations with different levels of subjectivity, such as an essay and a math test; and they are also similar across assessments made at different points of the year, so there is no kind of learning.

We also find another stage of the academic evaluation process where teachers impose different standards for students based on their classroom behavior. Conditional on end-of-year grade, teachers' decision to promote pupils below the passing cutoff grade is influenced by their behaviors. The well-behaved ones are in a higher proportion approved without having to take a reassessment examination. The opposite is true for the ill-behaved, which may have a 20% higher chance of going through a reassessment phase instead of being approved directly by teachers. These results are robust to unobservable student characteristics that could also influence teachers' decisions. Finally, in counterfactual exercises, we find that teacher assessment practices may significantly change the retention probability among students with different classroom behaviors.

Taken together, our results are relevant to the understanding of educational gaps between students with different behavioral characteristics. They indicate that teachers' assessment practices may impose or reduce obstacles to students' acquisition of skills and educational credentials depending on how they behave in class. From a policy standpoint, these results go against the objectives sought by recent reforms in the United States that aim to improve school evaluation systems, making them more standardized and fair (Feldman, 2018; McMillan, 2013). One of the measures proposed by these reforms is the complete separation of behavior and achievement assessments. In our context, these are also the guidelines. There are formative assessments of behavior and subjective ratings of overall student behavior. Still, teachers factor in their pupils' classroom behaviors into high-stakes decisions taken during the whole academic year in such a way that is not consistent with official guidelines. Hence, it may be privately optimal for them to use off-the-book channels to praise well-behaved pupils or punish those with poor behavior. These results highlight the challenge of encouraging teachers to be completely objective when giving an assessment. We left for future work to design and evaluate interventions aimed at reducing the influence of student behaviors on teachers' grading while not encouraging misbehavior.

References

Alan, S., Boneva, T., and Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, 134(3):1121–1162.

Alan, S. and Ertac, S. (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy*, 126(5):1865–1911.

Alesina, A., Carlana, M., Ferrara, E. L., and Pinotti, P. (2018). Revealing stereotypes: Evidence from immigrants in schools. Technical report, National Bureau of Economic Research.

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, volume 4, pages 1–181. Elsevier.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184.

Altonji, J. G. and Pierret, C. R. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Bertrand, M. and Pan, J. (2013). The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, 5(1):32–64.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436.

Bond, T. N. and Lang, K. (2018). The black–white education scaled test-score gap in grades k-7. *Journal of Human Resources*, 53(4):891–917.

Borghans, L., Golsteyn, B. H., Heckman, J. J., and Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences*, 113(47):13354–13359.

Borghans, L., Meijers, H., and Ter Weel, B. (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic inquiry*, 46(1):2–12.

Botelho, F., Madeira, R. A., and Rangel, M. A. (2015). Racial discrimination in grading: Evidence from brazil. *American Economic Journal: Applied Economics*, 7(4):37–52.

Breda, T. and Ly, S. T. (2015). Professors in core science fields are not always biased against women: Evidence from france. *American Economic Journal: Applied Economics*, 7(4):53–75.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied measurement in education*, 10(2):161–180.

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

Castillo, M., Ferraro, P. J., Jordan, J. L., and Petrie, R. (2011). The today and tomorrow of kids: Time preferences and educational outcomes of children. *Journal of Public Economics*, 95(11-12):1377–1385.

Cornwell, C., Mustard, D. B., and Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human resources*, 48(1):236–264.

Credé, M. and Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on psychological science*, 3(6):425–453.

Cubel, M., Nuevo-Chiquero, A., Sanchez-Pages, S., and Vidal-Fernandez, M. (2016). Do personality traits affect productivity? evidence from the laboratory. *The Economic Journal*, 126(592):654–681.

De Paola, M. and Gioia, F. (2017). Impatience and academic performance. less effort and less ambitious goals. *Journal of Policy Modeling*, 39(3):443–460.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. (2019). The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics*, 11(3):382–423.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, 132(4):1593–1640.

Diamond, R. and Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. Technical report, National Bureau of Economic Research.

Duckworth, A. L. and Allred, K. M. (2012). Temperament in the classroom.

Duckworth, A. L., Quinn, P. D., and Tsukayama, E. (2012). What no child left behind leaves behind: The roles of iq and self-control in predicting standardized achievement test scores and report card grades. *Journal of educational psychology*, 104(2):439.

Duckworth, A. L. and Seligman, M. E. (2005). Self-discipline outdoes iq in predicting academic performance of adolescents. *Psychological science*, 16(12):939–944.

Eren, O., Lovenheim, M. F., and Mocan, N. H. (2018). The effect of grade retention on adult crime: Evidence from a test-based promotion policy. Technical report, National Bureau of Economic Research.

Falch, T. and Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review*, 36:12–25.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., and Beechum, N. O. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance–A Critical Literature Review.* ERIC.

Feldman, J. (2018). School grading policies are failing children: A call to action for equitable grading. *Oakland, CA: Crescendo Education Group.*

Ferman, B. (2019). A simple way to assess inference methods. *arXiv preprint arXiv:1912.08772.*

Golsteyn, B., Non, A., and Zölitz, U. (2021). The impact of peer personality on academic achievement. *Journal of Political Economy*, Forthcoming.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–68.

Harlen, W. (2005). Teachers' summative practices and assessment for learning–tensions and synergies. *Curriculum Journal*, 16(2):207–223.

Heckman, J. J., Humphries, J. E., Veramendi, G., and Urzua, S. S. (2014). Education, health and wages. Technical report, National Bureau of Economic Research.

Heckman, J. J., Jagelka, T., and Kautz, T. D. (2019). Some contributions of economics to the study of personality. Technical report, National Bureau of Economic Research.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour economics*, 19(4):451–464.

Heckman, J. J., Pinto, R., and Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–86.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, 24(3):411–482.

Hinnerich, B. T., Höglin, E., and Johannesson, M. (2011). Are boys discriminated in swedish high schools? *Economics of Education review*, 30(4):682–690.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., and Kiguel, S. (2020). School effects on socio-emotional development, school-based arrests, and educational attainment. Technical report, National Bureau of Economic Research.

Johnson, R. L., Green, S. K., Kim, D.-H., and Pope, N. S. (2008). Educational leaders' perceptions about ethical practices in student evaluation. *American Journal of Evaluation*, 29(4):520–530.

Kautz, T. and Zanoni, W. (2014). *Measuring and fostering non-cognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal Program*. University of Chicago Chicago, IL.

Keogh, B. K. (1986). Temperament and schooling: meaning of" goodness of fit"? *New directions for child development*.

Khwaja, A. I., Andrabi, T., Das, J., Zajonc, T., et al. (2011). Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal: Applied Economics*.

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., and Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment*, 30(2):148–159.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4):241–253.

Lavecchia, A. M., Liu, H., and Oreopoulos, P. (2016). Behavioral economics of education: Progress and possibilities. In *Handbook of the Economics of Education*, volume 5, pages 1–74. Elsevier.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Political Economy*, 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. Technical report, 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. (2018). Charity begins at home (and at school): Effects of religion-based discrimination in education. Technical report, National Bureau of Economic Research.

Lockwood, J. and McCaffrey, D. F. (2014). Correcting for test score measurement error in ancova models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, 39(1):22–52.

Lubbers, M. J., Van Der Werf, M. P., Kuyper, H., and Hendriks, A. J. (2010). Does homework behavior mediate the relation between personality and academic performance? *Learning and Individual differences*, 20(3):203–208.

Manacorda, M. (2012). The cost of grade retention. *Review of Economics and Statistics*, 94(2):596–606.

McMillan, J. H. (2013). *SAGE handbook of research on classroom assessment*. Sage.

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *The review of economic studies*, 76(4):1431–1459.

Non, A. and Tempelaar, D. (2016). Time preferences, study effort, and academic performance. *Economics of Education Review*, 54:36–61.

Nordin, M., Heckley, G., and Gerdtham, U. (2019). The impact of grade inflation on higher education enrolment and earnings. *Economics of Education Review*, 73:101936.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Papageorge, N. W., Ronda, V., and Zheng, Y. (2019). The economic value of breaking bad: Misbehavior, schooling and the labor market. Technical report, National Bureau of Economic Research.

Piché, G. L., Michlin, M., Rubin, D., and Sullivan, A. (1977). Effects of dialect-ethnicity, social class and quality of written compositions on teachers' subjective evaluations of children. *Communications Monographs*, 44(1):60–72.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on timss. *Applied Measurement in Education*, 17(1):1–24.

Roen, D. (1992). Gender and teacher response to student writing. In *Gender issues in the teaching of English*, pages 126–141. Heinemann.

Segal, C. (2008). Classroom behavior. *Journal of Human Resources*, 43(4):783–814.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, 58(8):1438–1457.

Segal, C. (2013). Misbehavior, education, and labor market outcomes. *Journal of the European Economic Association*, 11(4):743–779.

Spengler, M., Damian, R. I., and Roberts, B. W. (2018). How you behave in school predicts life success above and beyond family background, broad traits, and cognitive ability. *Journal of Personality and Social Psychology*, 114(4):620.

Starch, D. and Elliott, E. (1912). Reliability of grading high-school work in english. *The School Review*, 20(7):442–457.

Starch, D. and Elliott, E. C. (1913). Reliability of grading work in mathematics. *The School Review*, 21(4):254–259.

Sutter, M., Kocher, M. G., Glätzle-Rützler, D., and Trautmann, S. T. (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review*, 103(1):510–31.

Terrier, C. (2016). Boys lag behind: How teachers' gender biases affect student achievement. *IZA Discussion Paper*.

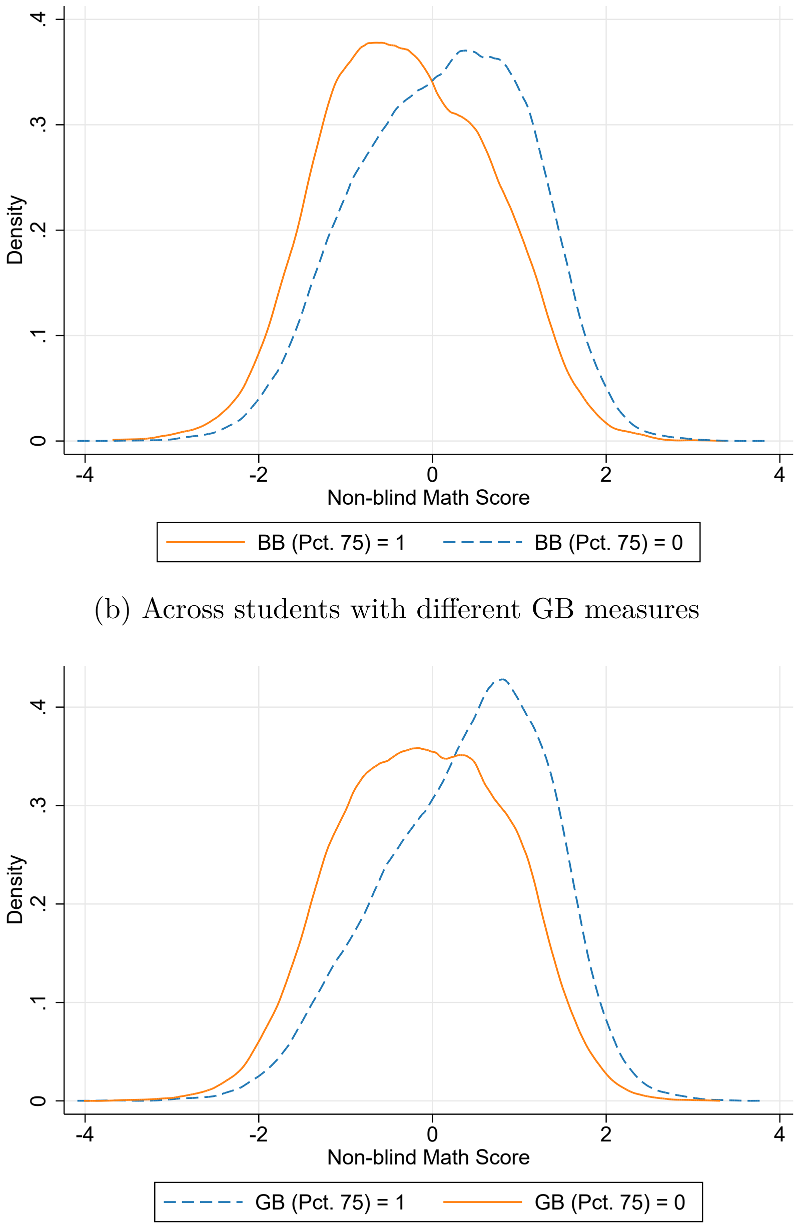
Wen, S.-s. (1979). Racial halo on evaluative rating: General or differential? *Contemporary Educational Psychology*, 4(1):15–19.

Wyatt-Smith, C., Klenowski, V., and Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, policy & practice*, 17(1):59–75.

Xu, J. and Gong, J. (2017). Statistical discrimination, taste-based bias, and cognitive bias – analyzing grading bias caused by handwriting quality in a randomized control trial. [Available as https://www.dropbox.com/s/orb20ml2zkt5rpu/JMP_handwriting_Jianfeng_Xu.pdf?dl=0. Accessed 20-July-2020].

Tables and Figures

Figure (1) Distribution of Non-blind Math Scores

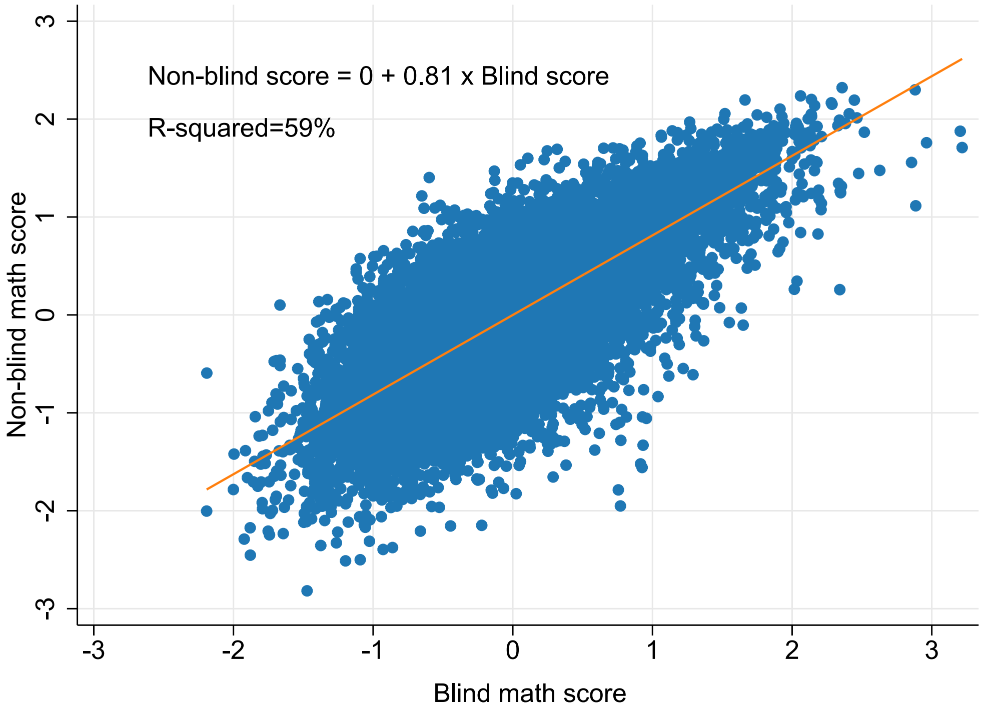


(a) Across students with different BB measures

(b) Across students with different GB measures

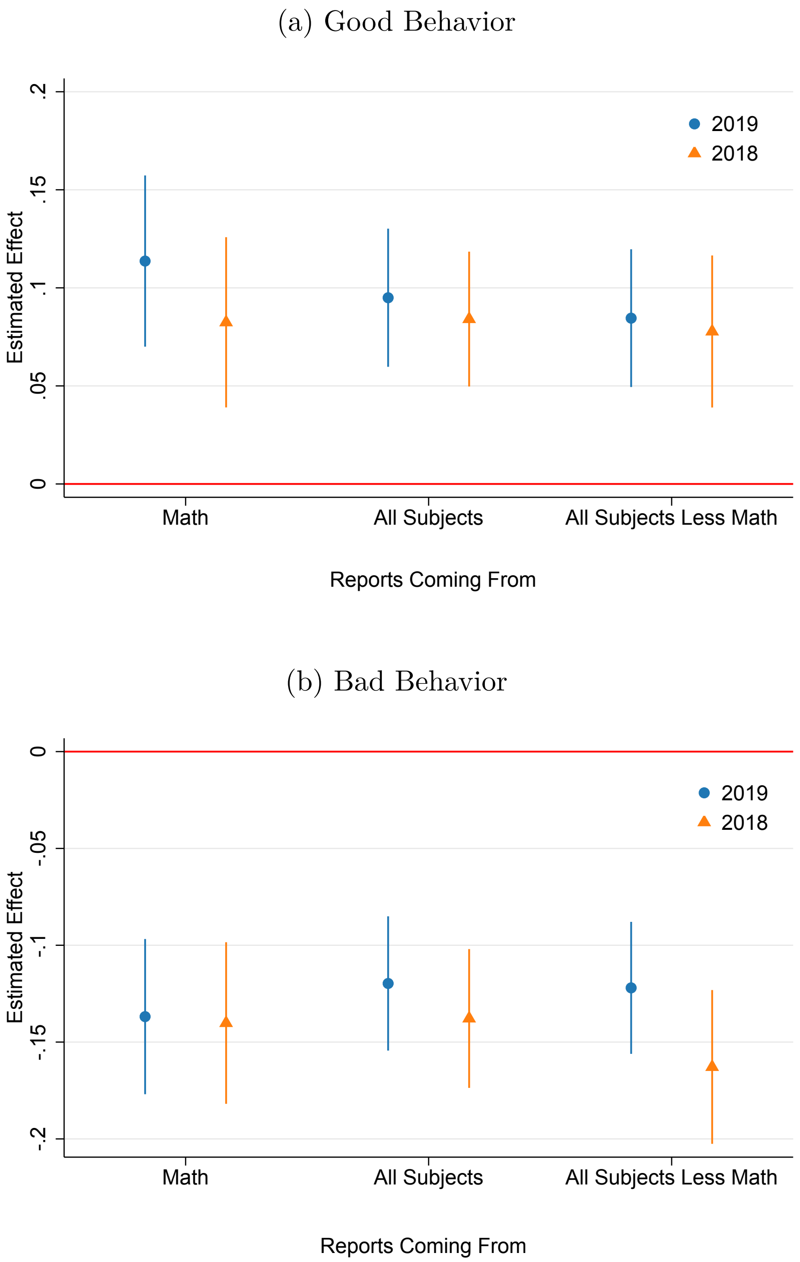
Note: These figures plot the distribution of non-blind math scores. Observations are at the student×exam level. $BB(Pct.75)$ and $GB(Pct.75)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In panel (a), solid line represents students with $BB(Pct.75) = 1$ and the dotted line represents those with $BB(Pct.75) = 0$. In panel (b), solid line represents students with $GB(Pct.75) = 0$ and the dotted line represents those with $GB(Pct.75) = 1$. All test scores are standardized (the mean equals zero and the variance equals one).

Figure (2) Association Between the Blind and Non-blind Math Scores



Note: This figure plots the average of the blind and non-blind math scores across all examinations. The line fits the data points by OLS. All test scores are standardized (the mean equals zero and the variance equals one).

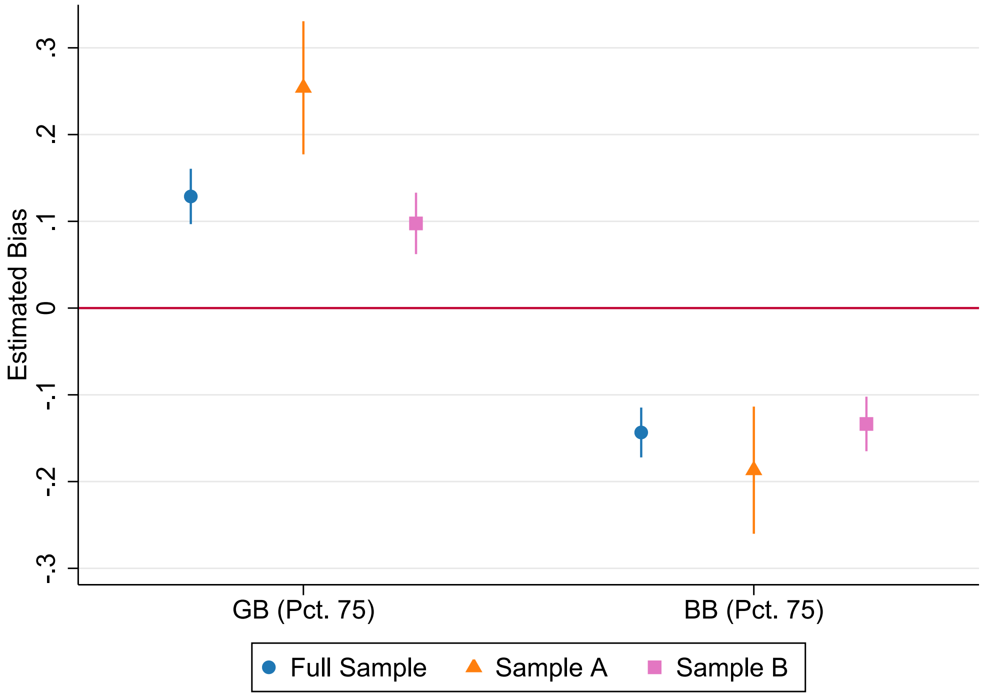
Figure (3) Estimated biases in the non-blind scores toward classroom behavior while using different ways of measuring behavior


(a) Good Behavior

(b) Bad Behavior

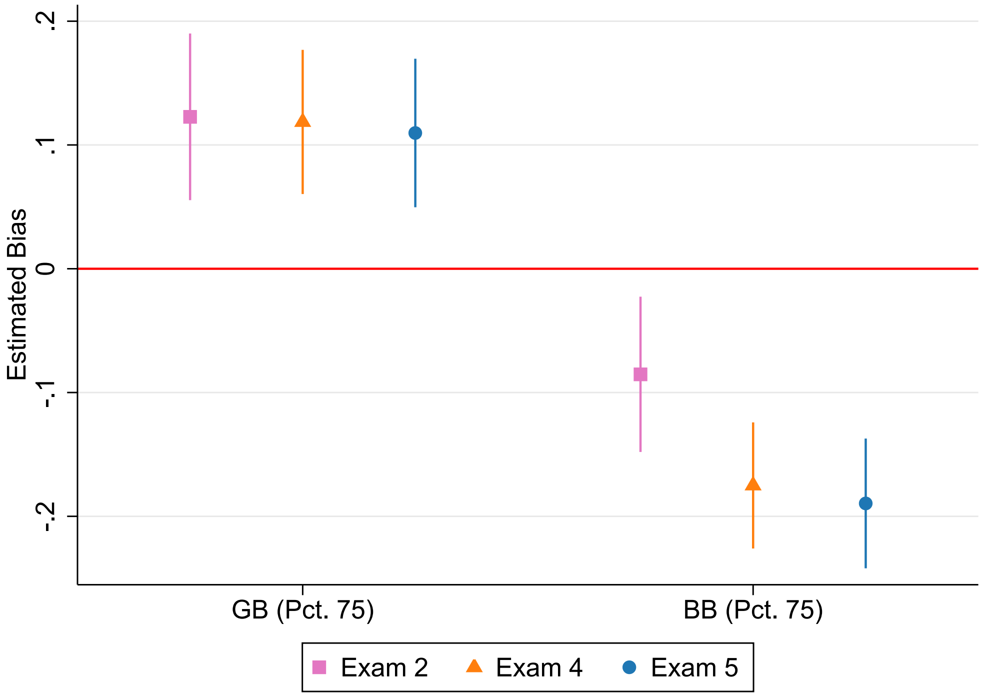
Note: This figure plots 95% confidence intervals computed with student-level cluster and point estimates from student×exam-level IV regressions of teacher-assigned math scores on classroom behavior, using different behavior measures. Panels a and b report the estimated effects for GB(Pct. 75) and BB(Pct. 75), respectively, which stand for binary variables that indicate whether students are at the top quartile of the behavior measures' distribution. Results vary over whether measures use the reports coming from math teachers, teachers from all subjects, or teachers from all subjects expect math; and whether they use the reports from current teachers (2019) or past teachers (2018). All specifications follow Table 2, column 4.

Figure (4) Estimated biases in classrooms where students with different in-class behaviors differ (Sample B) and do not differ (Sample A) in their math proficiency



Note: This figure plots 95% confidence intervals computed with student-level cluster and point estimates from student×exam-level IV regressions of teacher-assigned math scores from the second semester on classroom behavior (measures use the reports from the first semester only), for different samples. Sample A selects classrooms where in the first semester students with $BB(pct.75) = 1$ or $GB(pct.75) = 0$ performed, on average, better in the blind math exams than their classmates with $BB(pct.75) = 0$ or $GB(pct.75) = 1$, respectively. The subsample where the previous conditions are not satisfied is called of sample B. Full sample uses samples A and B. All specifications follow Table 2, column 4.

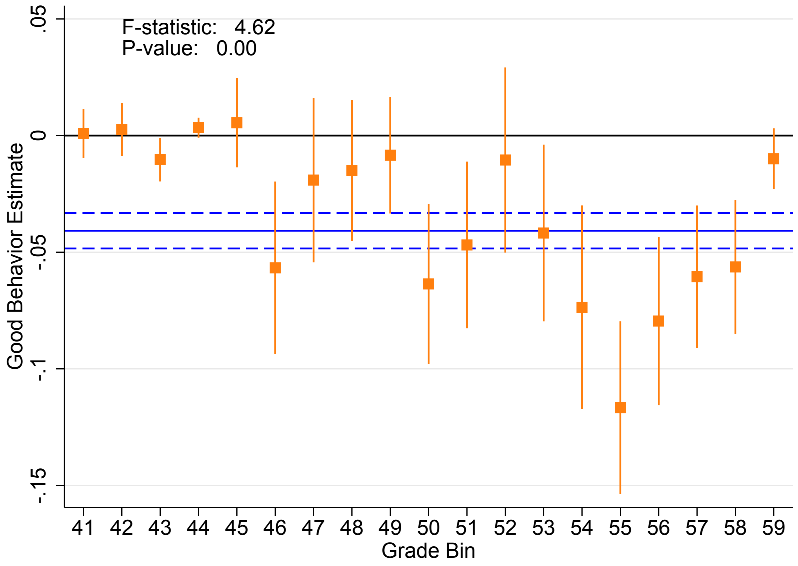
Figure (5) Estimated biases in the non-blind scores toward classroom behavior across the year



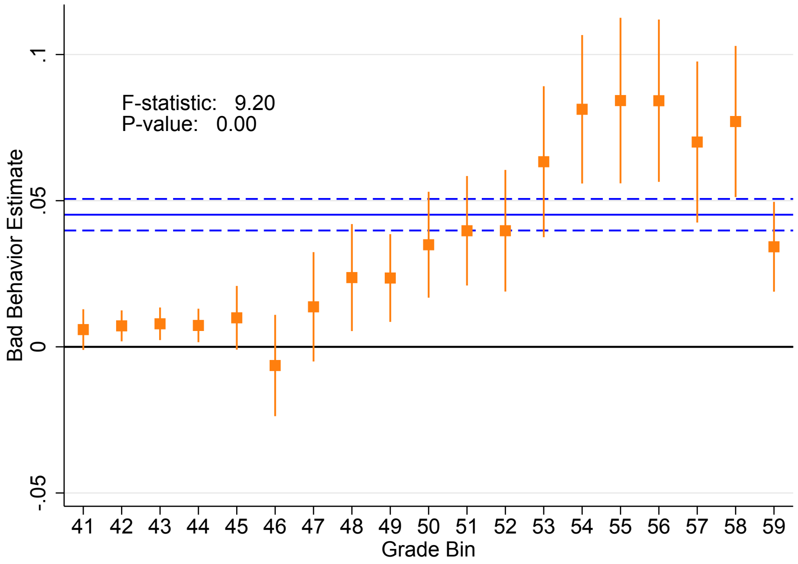
Note: This figure plots 95% confidence intervals computed with student-level cluster and point estimates from student×exam-level IV regressions of teacher-assigned math scores on classroom behavior, using subsamples that are specific for each exam. $BB(Pct.75)$ and $GB(Pct.75)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. All specifications follow Table 2, column 4.

Figure (6) Differential probability of going through the reassessment phase across the final grade, by behavioral group

(a) Good Behavior

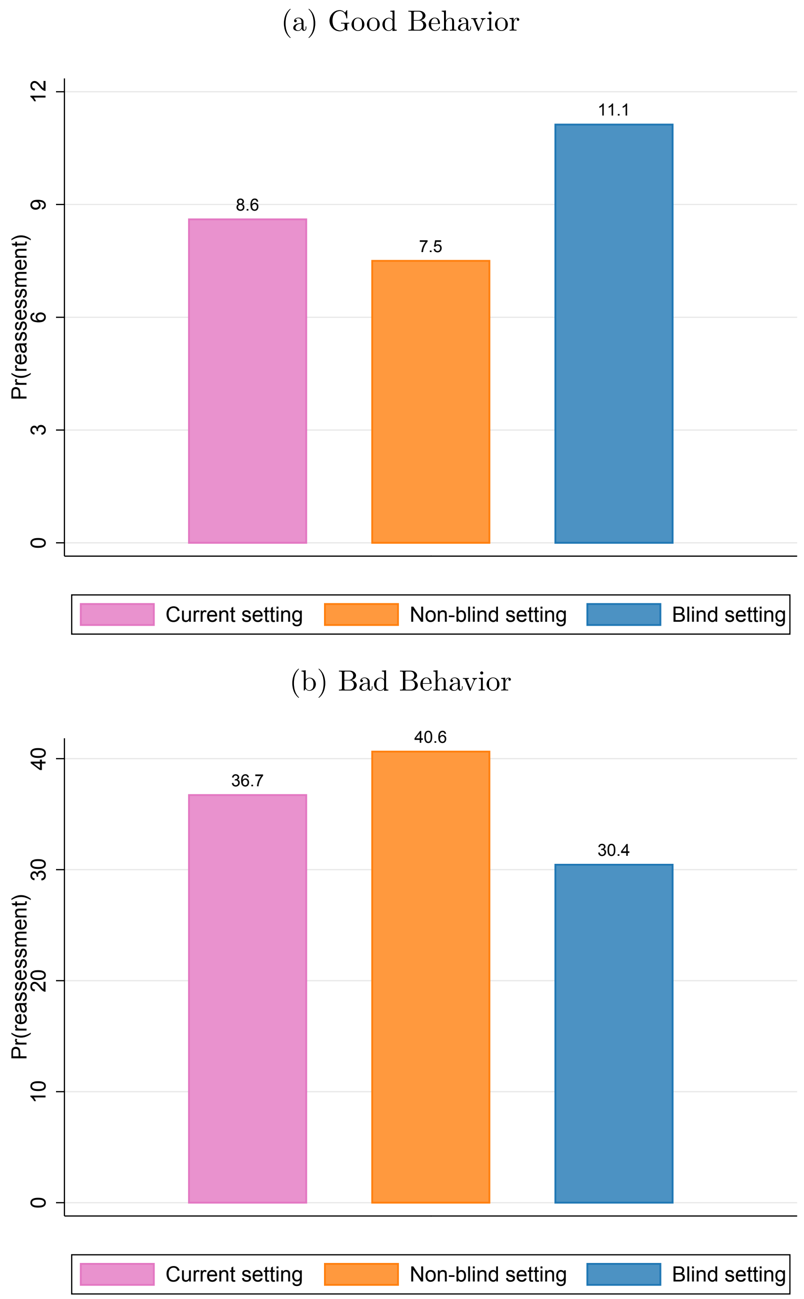


(b) Bad Behavior



Note: This figure estimates student i × subject s-level OLS regressions of a dummy indicating whether i had to go through the reassessment phase in subject s on classroom behavior and additional controls, for each final grade bin (of size 1) between 41 and 59. All specifications include controls for age, ethnicity, gender, and fixed effects for school, grade, and subject. Panels a and b report the estimated effects for GB(Pct. 75) and BB(Pct. 75), respectively, which stand for binary variables that indicate whether students are at the top quartile of the behavior measures' distribution. These measures are computed using the assessments from all teachers. Confidence intervals (95%) are computed with student-level clusters. The solid blue line represents the average effect computed by a bin-size-weighted average of the grade-bin effects. The dashed blue lines represent respective 95% confidence intervals.

Figure (7) Probability of going through the reassessment phase under different counterfactual scenarios

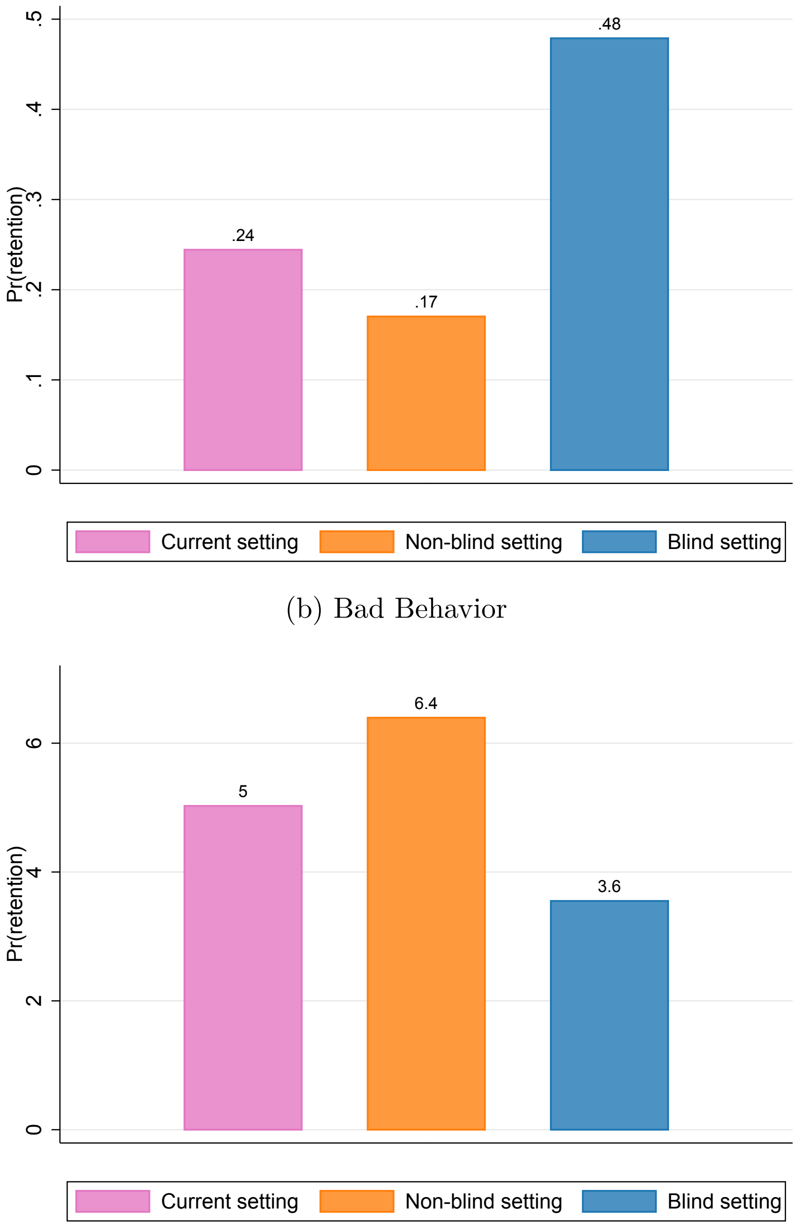


(a) Good Behavior

(b) Bad Behavior

Note: This figure simulates the proportion of students with good (panel a) and bad (panel b) behavior – GB(pct.75) and BB(pct.75) – going through the reassessment phase under different counterfactual scenarios. "Current setting" uses the original setting. "Non-blind setting" includes teacher grading biases toward behavior in the blindly-assigned scores. "Blind setting" eliminates the influence of classroom behavior into grading and approval decisions. Counterfactual grading scenarios are computed using the estimates from Section 3 and 5.

Figure (8) Probability of grade retention under different counterfactual scenarios



(a) Good Behavior

(b) Bad Behavior

Note: This figures simulates the grade repetition rate among students with good (panel a) and bad (panel b) behavior – GB(pct.75) and BB(pct.75) – under different counterfactual scenarios. “Current setting” uses the original setting. “Non-blind setting” includes teacher grading biases toward behavior in the blindly-assigned scores. “Blind setting” eliminates the influence of classroom behavior into grading and approval decisions. Counterfactual grading scenarios are computed using the estimates from Section 3 and 5.

Table (1) Summary Statistics

Variable	Mean	Observations
Students		23001
Schools		80
Classrooms		738
Grade		
Six	0.15	3446
Seven	0.16	3600
Eight	0.16	3708
Nine	0.15	3486
Ten	0.19	4342
Eleven	0.19	4419
Demographics		
Age	14.16	23001
Over-age	0.05	23001
Girl	0.53	12296
White	0.58	13344
Brown	0.24	5603
Non-declared	0.15	3338
Black	0.03	597
Other (Yellow or Indigenous)	0.01	123
Behavior Data (all reports)		
Classes with at least one good report	0.99	
Classes with at least one bad report	0.98	
Good reports	44.18	22187
Bad reports	11.51	22187
Good behavior measure	0.53	22186
Bad behavior measure	0.29	22020
Behavior Data (math reports)		
Classes with at least one good report	0.84	
Classes with at least one bad report	0.81	
Good reports	10.20	17115
Bad reports	3.22	17115
Good behavior measure	0.47	17048
Bad behavior measure	0.25	16389

Note: This table reports summary statistics for our data. Except for "Schools" and "Classrooms", observations refer to the number of students. Over-age indicates whether a student is two or more years older than the official age for a specific grade.

Table (2) Estimated biases in the non-blind math scores toward classroom behavior

	(1)	(2)	(3)	(4)	(5)
VARIABLES	OLS	IV	IV	IV	IV
GB (Pct. 75)	0.649	0.119	0.113	0.112	0.119
	(0.017)***	(0.017)***	(0.018)***	(0.017)***	(0.017)***
BB (Pct. 75)	-0.418	-0.159	-0.147	-0.139	-0.123
	(0.015)***	(0.016)***	(0.016)***	(0.015)***	(0.016)***
Blind Math Score		1.112	1.112	1.010	0.927
		(0.015)***	(0.016)***	(0.034)***	(0.103)***
Student Demographics	No	No	Yes	Yes	Yes
Other Scores	No	No	No	Yes	Yes
Instrumenting Language Scores	-	No	No	No	Yes
High-Order Polynomials for Scores	-	No	No	No	Yes
Number of Observations	36044	36044	36044	36044	36044
Number of Clusters	14692	14692	14692	14692	14692
First-stage F Statistic		4513	4343	1099	6.775

Note: This table reports student×exam-level OLS (column 1) and IV (columns 2-5) regressions of teacher-assigned math scores on classroom behavior. $BB(Pct.75)$ and $GB(Pct.75)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In the IV estimates, lagged blind math scores are used as instrumental variable for the current math scores. All specifications include classroom fixed effects and exams fixed effects. Other scores include the cumulative average performance in science and humanities, and current performance in language. High-order polynomials for scores include a third order polynomial for blind math scores, and an interaction term between math and language scores. In Column 5, we also use lagged language scores as instrumental variable for the current language scores. Controls include indicators for age, gender, and ethnicity (Black, Indigenous, *Pardo*, Yellow, and White). We also include a dummy for students with missing data on ethnicity. Standard errors in parenthesis are robust and clustered at the student level.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (3) Estimated biases in the non-blind math scores toward classroom behavior (OLS)

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	OLS	OLS	OLS	OLS	OLS	OLS - DID
GB (Pct. 75)	0.649	0.444	0.443	0.261	0.233	0.132
	(0.017)***	(0.014)***	(0.014)***	(0.012)***	(0.012)***	(0.015)***
BB (Pct. 75)	-0.418	-0.318	-0.321	-0.189	-0.171	-0.154
	(0.015)***	(0.013)***	(0.013)***	(0.012)***	(0.011)***	(0.014)***
Blind Math Score		0.429	0.423	0.210	0.161	
		(0.005)***	(0.006)***	(0.005)***	(0.005)***	
Student Demographics	No	No	Yes	Yes	Yes	Yes
Other Scores	No	No	No	Yes	Yes	Yes
Past Math Scores	No	No	No	No	Yes	-
Number of Observations	36044	36044	36044	36044	36044	36044
Number of Clusters	14692	14692	14692	14692	14692	14692
Adjusted R-squared	0.107	0.283	0.289	0.441	0.475	0.00735

Note: This table reports student×exam-level OLS regressions of teacher-assigned math scores on classroom behavior (columns 1-5). In column (6), the dependent variable is the difference between teacher-assigned and blindly-assigned math scores. $BB(Pct.75)$ and $GB(Pct.75)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. All specifications include classroom fixed effects and exams fixed effects. Other scores include the cumulative average performance in science and humanities, and current performance in language. Past math scores include the lagged blind math score. Controls for ethnicity include 5 indicators: Black, Indigenous, *Pardo*, Yellow, and White. We also include a dummy for students with missing data on ethnicity. Standard errors in parenthesis are robust and clustered at the student level.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (4) Estimated biases in the SEs' blind scores

	(1)	(2)
VARIABLES	OLS	IV
GB (Pct. 75)	0.497	0.033
	(0.022)***	(0.022)
BB (Pct. 75)	-0.211	-0.011
	(0.021)***	(0.021)
Blind Math Score		0.804
		(0.045)***
Number of Observations	16166	16166
Number of Clusters	10446	10446
First-stage F Statistic		633.5

Note: This table reports student×exam-level OLS (columns 1) and IV (columns 2) regressions of math test scores from blindly-graded specific exams on classroom behavior. Column 2 follows the same specification from Table 2, column 4.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

A Additional Figures and Tables

Figure (A.1) Schools' Setting

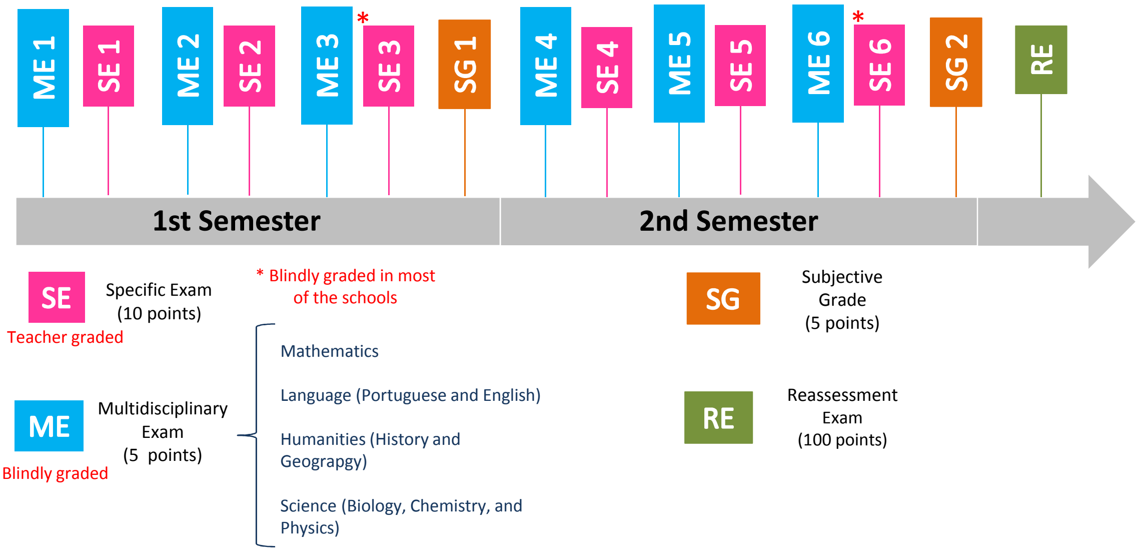
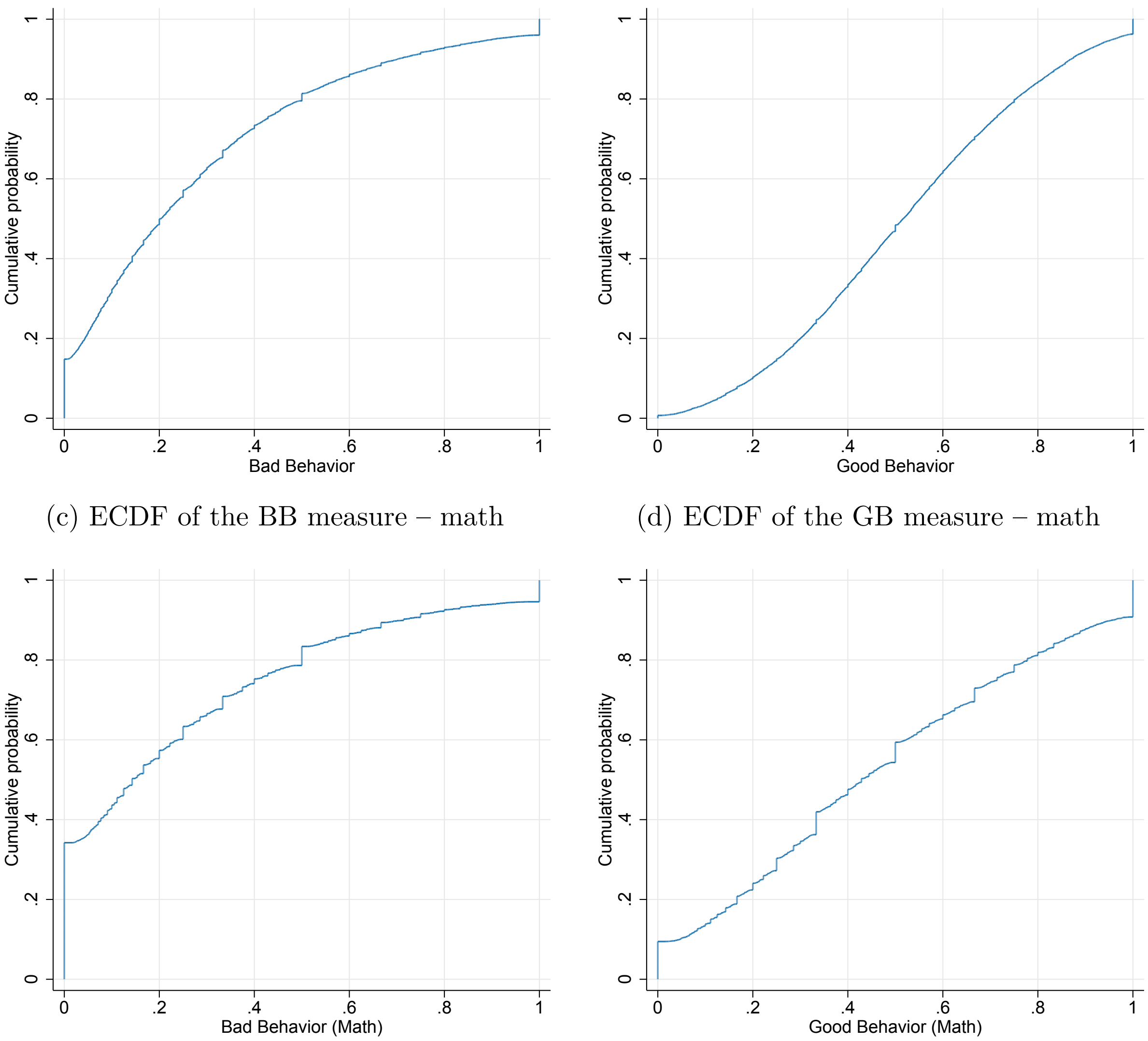

ME 1
SE 1
ME 2
SE 2
ME 3
*
SE 3
SG 1
ME 4
SE 4
ME 5
SE 5
ME 6
*
SE 6
SG 2
RE
1st Semester
2nd Semester
SE
Specific Exam (10 points)
Teacher graded
* Blindly graded in most of the schools
ME
Multidisciplinary Exam (5 points)
Blindly graded
Mathematics
Language (Portuguese and English)
Humanities (History and Geograpgy)
Science (Biology, Chemistry, and Physics)
SG
Subjective Grade (5 points)
RE
Reassessment Exam (100 points)

Figure (A.2) Empirical CDF of the Good Behavior (GB) and Bad Behavior (BB) Measures



(a) ECDF of the BB measure

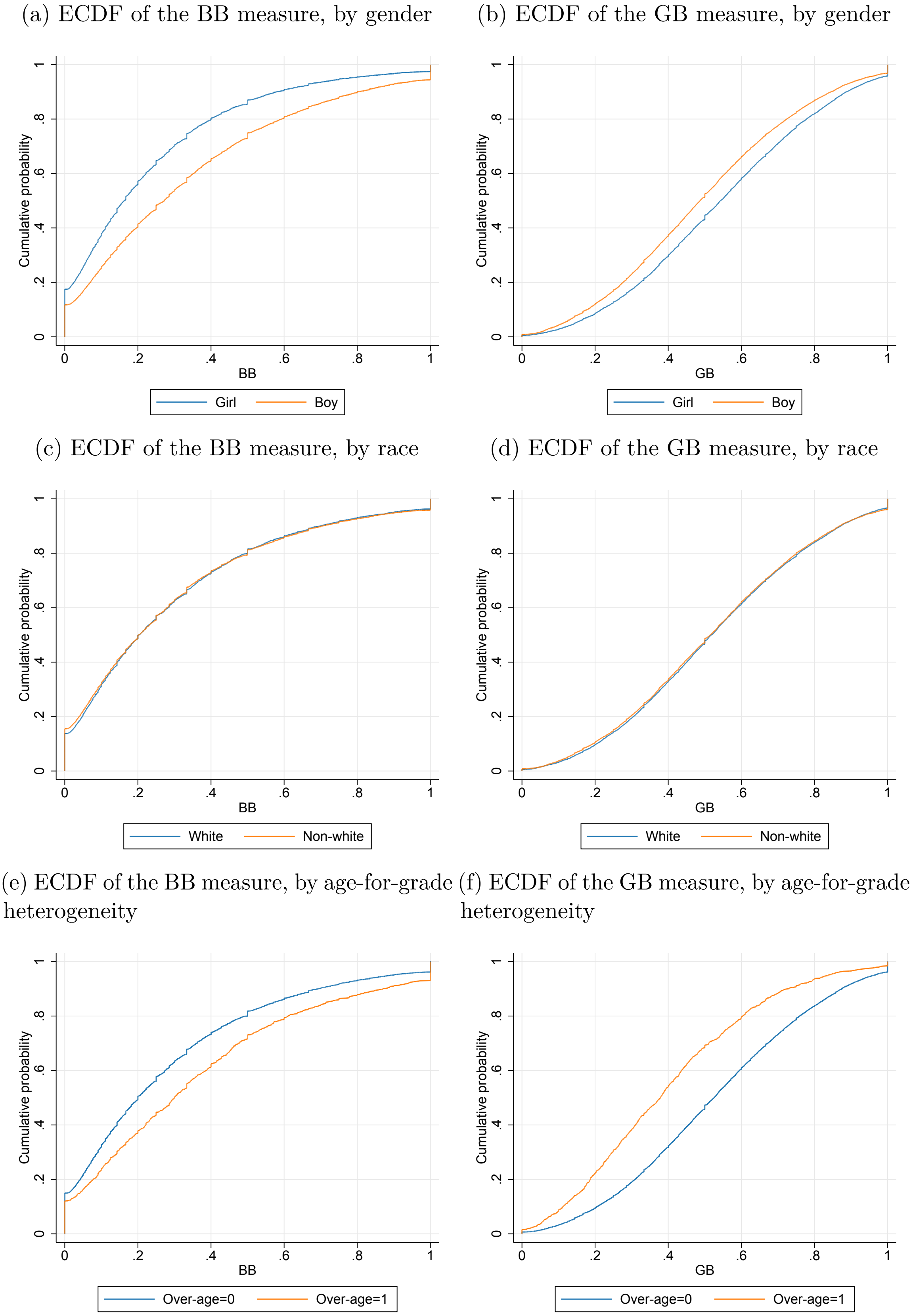
(b) ECDF of the GB measure

(c) ECDF of the BB measure – math

(d) ECDF of the GB measure – math

Note: This figure estimates the empirical cumulative distribution functions (ECDF) of the behavior measures. Panels a and b plot the ECDF of the bad and good behavior measures, respectively, computed using the assessments made by all teachers. Panels c and d plot the ECDF of the bad and good behavior measures, respectively, computed using only the assessments by math teachers.

Figure (A.3) Empirical CDF of the good behavior (GB) and bad behavior (BB) measures, by student demographics


(a) ECDF of the BB measure, by gender

(b) ECDF of the GB measure, by gender

(c) ECDF of the BB measure, by race

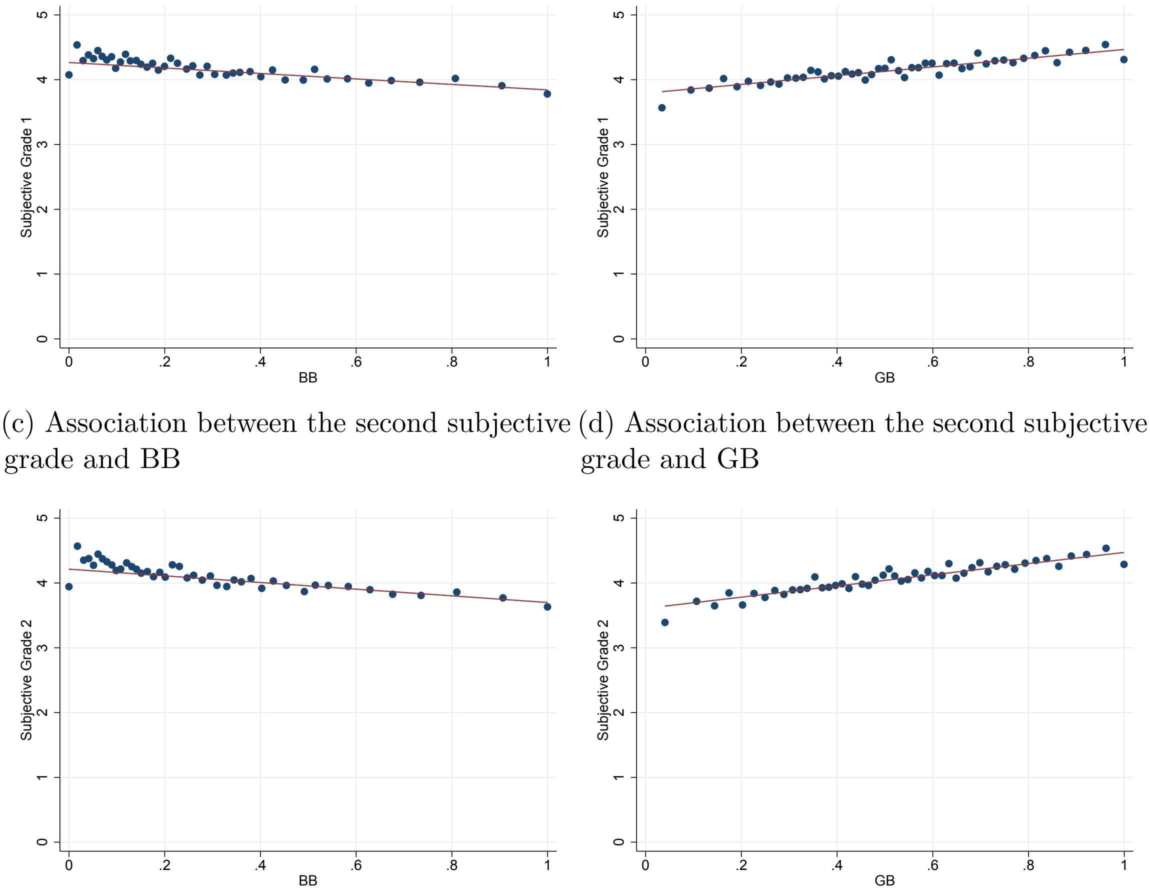
(d) ECDF of the GB measure, by race

(e) ECDF of the BB measure, by age-for-grade heterogeneity

(f) ECDF of the GB measure, by age-for-grade heterogeneity

Note: This figure estimates the empirical cumulative distribution functions (ECDF) of the behavior measures, by student demographics. Over-age indicates whether a student is two or more years older than the official age for a specific grade.

Figure (A.4) Mean relationship between the subjective grades and the behavior measures



(a) Association between the first subjective grade and BB

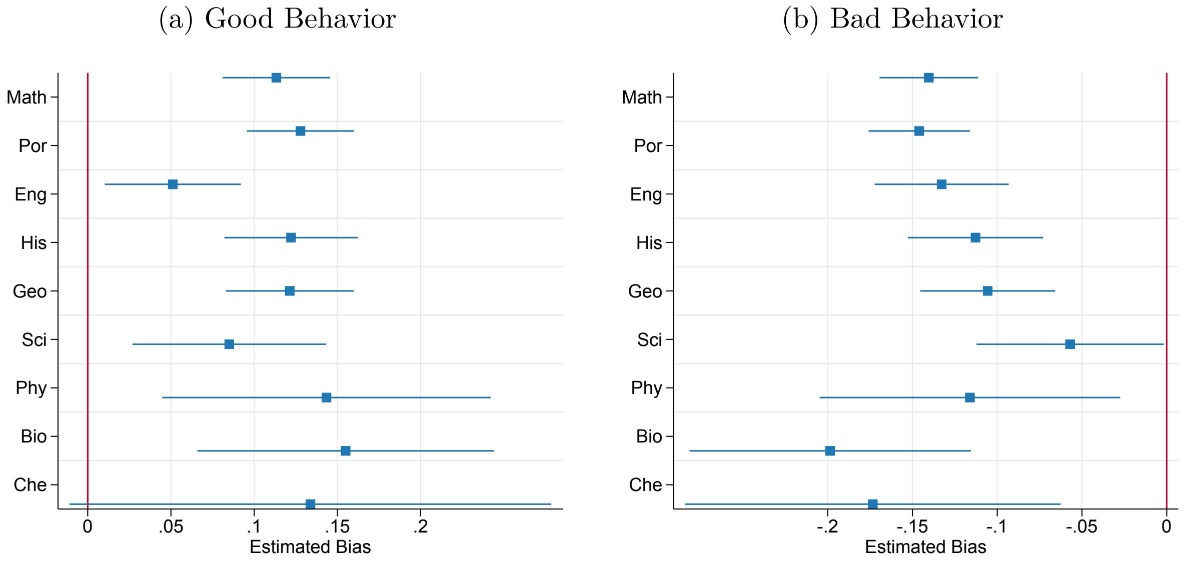
(b) Association between the first subjective grade and GB

(c) Association between the second subjective grade and BB

(d) Association between the second subjective grade and GB

Note: This figure plots binned scatterplots describing the mean relationship between the subjective grades and behavior measures.

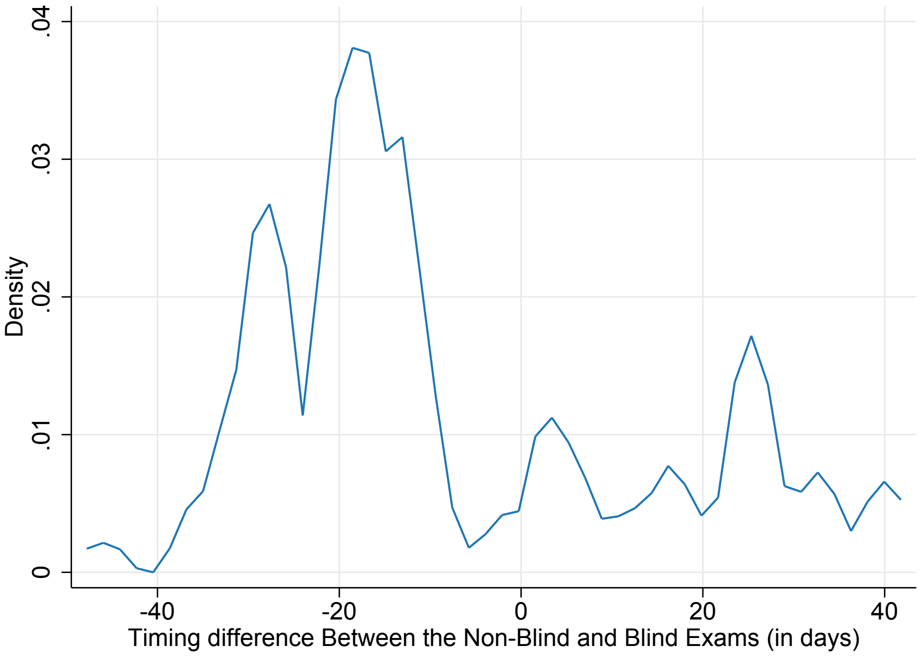
Figure (A.5) Estimated biases in the non-blind scores from Mathematics, Portuguese, English, History, Geography, Science, Physics, Biology, and Chemistry


(a) Good Behavior

(b) Bad Behavior

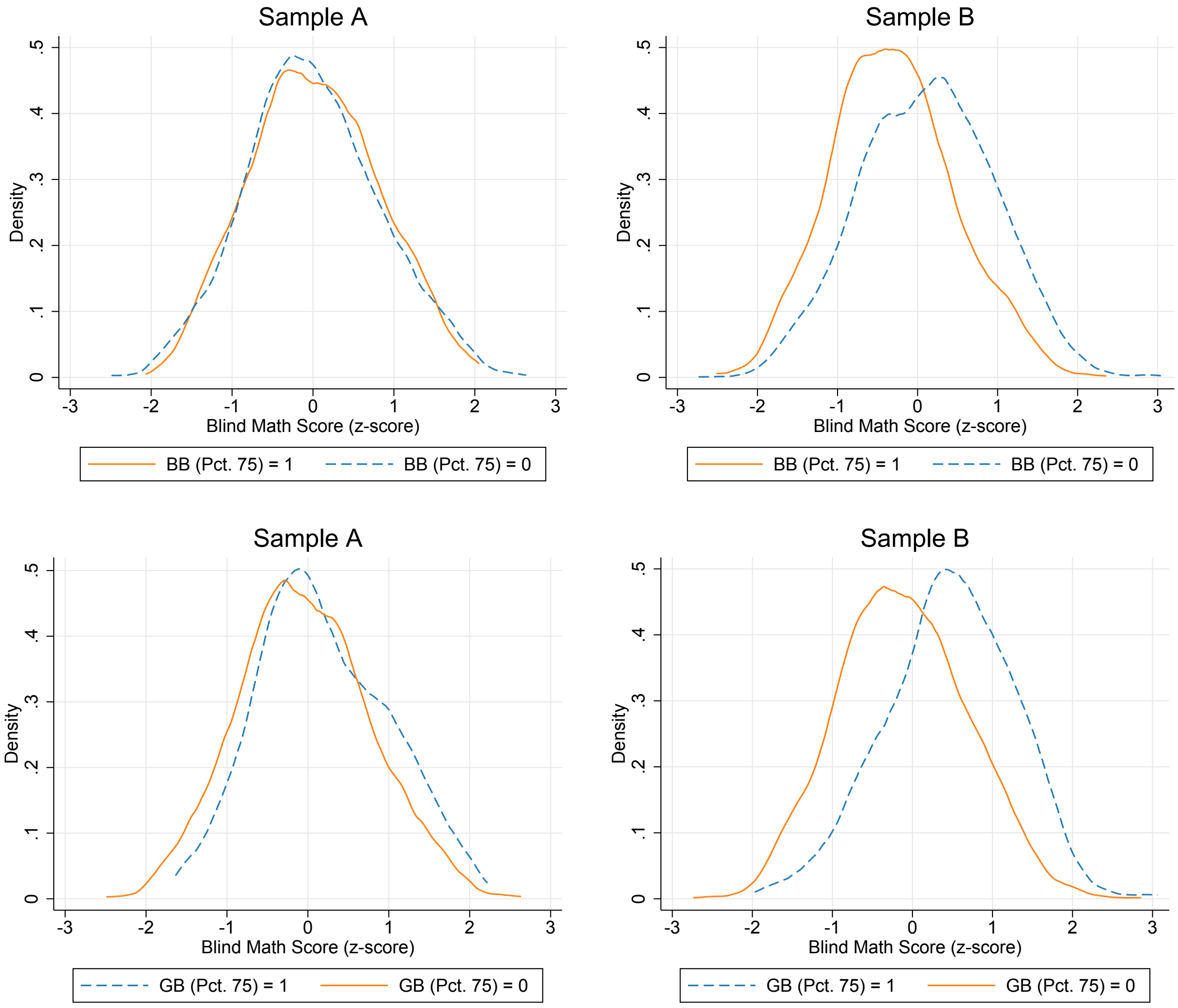
Note: This figure plots 95% confidence intervals computed with student-level cluster and point estimates from student×exam-level IV regressions of teacher-assigned scores from several subjects on classroom behavior, and proficiency in the material covered by the examination (measured by the blind scores,which we instrument using lagged blind scores). Panels a and b report the estimated effects for GB(Pct. 75) and BB(Pct. 75), respectively, which stand for binary variables that indicate whether students are at the top quartile of the behavior measures' distribution. When the dependent variable is a math test score, we measure proficiency using the blind math scores, and additionally control for past performance in blindly-graded science and language exams. When the dependent variable is a Portuguese or English test score, we measure proficiency using blind language score, and additionally control for past performance in blindly-graded math and science exams. When the dependent variable is a teacher-assigned score in science, chemistry, physics, or biology, we measure proficiency using blind science score, and additionally control for past performance in blindly-graded math and language exams. All specifications additionally include controls for age, gender, ethnic indicators and fixed effects for classroom and exam. Students from grades 6 through 9 do not have classes in chemistry, physics, and biology; only a science class. The opposite is true for students from grades 10 through 11.

Figure (A.6) Density of the timing differences between the blind and the non-blind exams



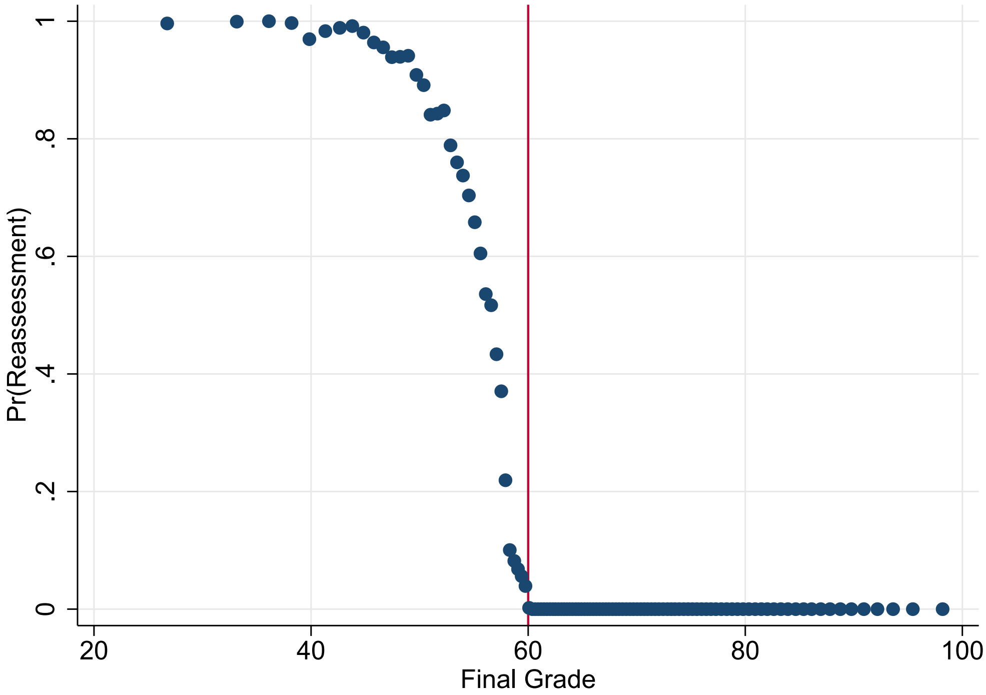
Note: This figure estimates the density of the timing differences between the blind and non-blind math exams, when we pool all the six exams.

Figure (A.7) Distribution of Blind Math Scores from the 1st Semester



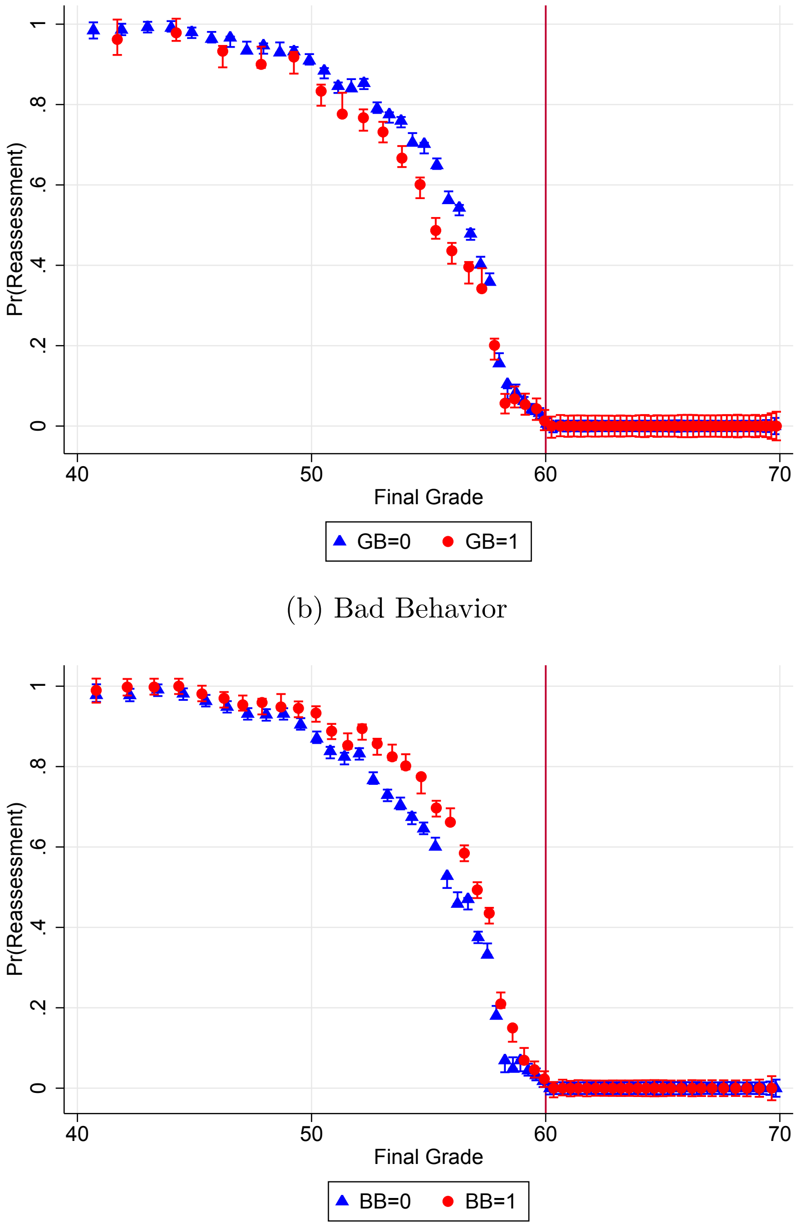
Note: These figures estimate the density of the blind math scores from the first-semester exams for students whose behavior indicators assume different values, using two different samples. $BB(Pct.75)$ and $GB(Pct.75)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In the top two figures, solid line represents students with $BB(Pct.75) = 1$ and the dotted line represents those with $BB(Pct.75) = 0$. In the bottom two figures, solid line represents students with $GB(Pct.75) = 0$ and the dotted line represents those with $GB(Pct.75) = 1$. Sample A selects classrooms where in the first semester students with $BB(pct.75) = 1$ or $GB(pct.75) = 0$ performed, on average, better in the blind math exams than their classmates with $BB(pct.75) = 0$ or $GB(pct.75) = 1$, respectively. We call this subsample of sample A. The subsample where the previous conditions are not satisfied is called of sample B.

Figure (A.8) Probability of going through the reassessment phase conditional on final grade



Note: This figure plots the proportion of students going through the reassessment phase across several final grade bins. In our data, all students that do not go through the reassessment phase are promoted to the next year.

Figure (A.9) Probability of going through the reassessment phase conditional on end-of-year grade, by in-class behavior group

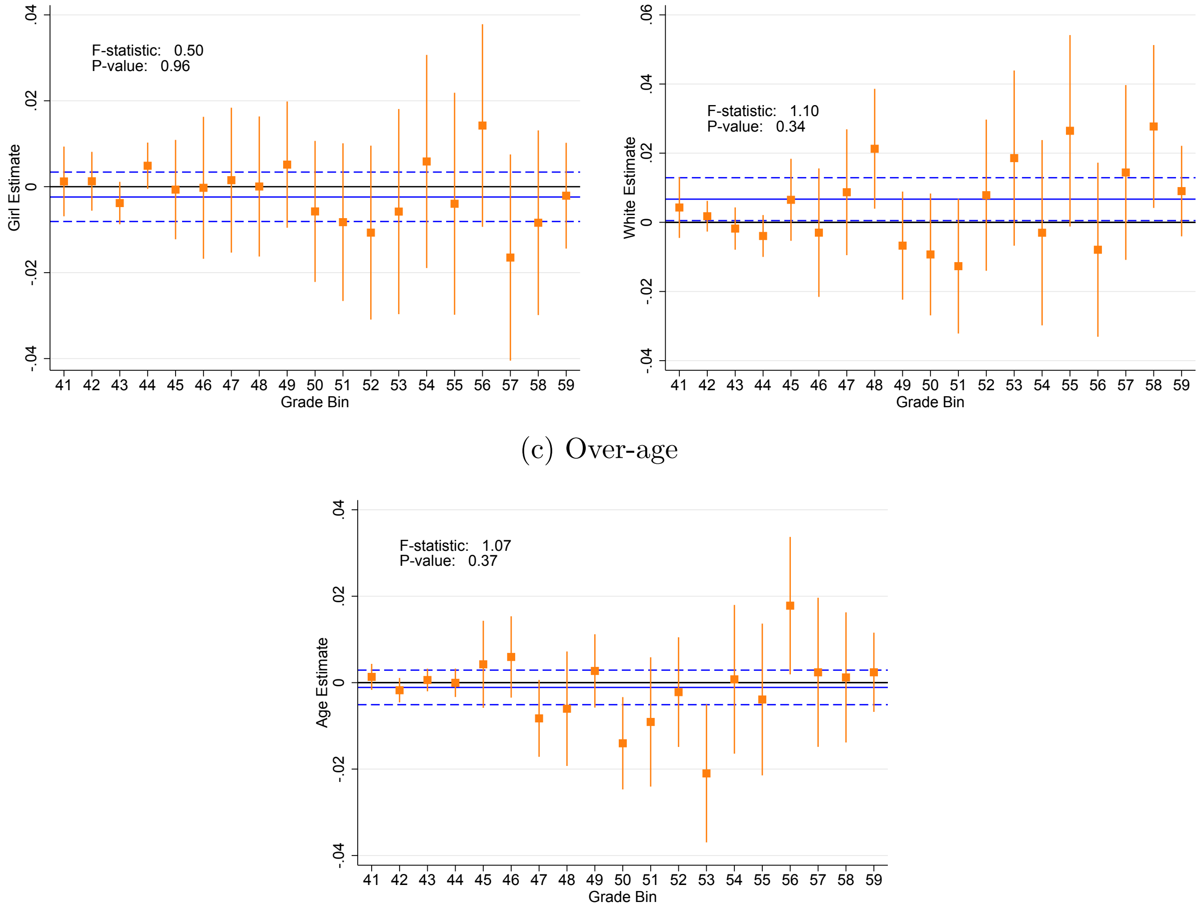


(a) Good Behavior

(b) Bad Behavior

Note: This figure plots the proportion of students going through the reassessment phase across several final grade bins, by different classroom behavior characteristics. $GB = 1$ ($BB = 1$) for students in the top quartile of the good (bad) behavior measure distributions. In our data, all students that do not go thought he reassessment phase are promoted to the next year.

Figure (A.10) Differential probability of going through the reassessment phase across end-of-year grade bins, by demographics

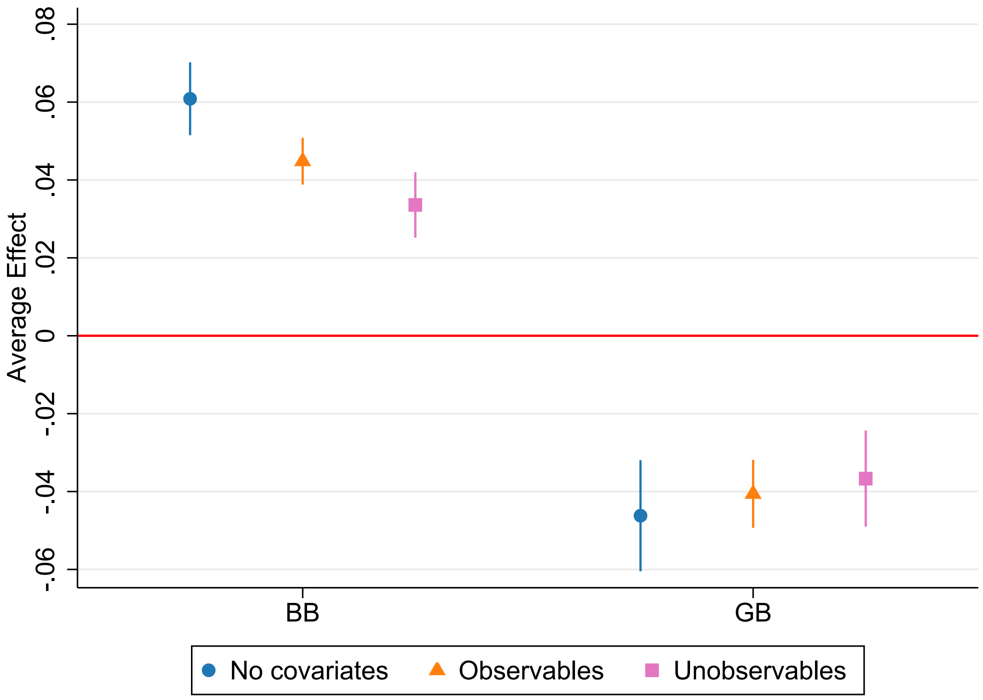

(a) Girl

(b) White

(c) Over-age

Note: This figure estimates student i × subject s-level OLS regressions of a dummy indicating whether i had to go through the reassessment phase in subject s on student demographics and additional controls, for each final grade bin (of size 1) between 41 and 59. All specifications include controls for classroom behavior and fixed effects for school, grade, and subject. Panels a, b, and c report the estimated effects for binary variables indicating whether i is girl, white, or in age-grade distortion. Confidence intervals (95%) are computed with student-level clusters. The solid blue line represents the average effect computed by a bin-size-weighted average of the grade-bin effects. The dashed blue lines represent respective 95% confidence intervals. This figure also presents the F-statistic (and respective p-value) from a joint hypothesis test under the null that all grade-bin estimates equal zero.

Figure (A.11) Stability of the results to observable and unobserved heterogeneity



Note: This figure uses the estimates from student i ×subject s-level OLS regressions of a dummy indicating whether i had to go through the reassessment phase in subject s on end-of-year grade bin fixed effect, classroom behavior indicators, additional controls (indicators for ethnicity, gender, and age-grade distortion), and fixed effects for school, grade, and subject, and the interaction between end-of-year grade bin fixed effects with the other variables from the model. "No covariates" plots the average effect (weighted by end-of-year grade bin size) of behavior indicators across the end-of-year grade bins in a model with no additional controls and fixed effects. "Observables" plots such an aggregation from a model with additional controls and fixed effects. "Unobservables" plots bias-adjusted treatment effects following Oster (2019) and her rule of thumbs for parametrization. 95% confidence intervals are computed using a bootstrap. BB and GB stand for binary variables that indicate whether students are at the top quartile of the behavior measures' distribution.

Table (B1) Estimated biases in the non-blind math scores toward classroom behavior – school-level cluster

	(1)	(2)
VARIABLES	OLS	IV
GB (Pct. 75)	0.649	0.112
	(0.024)***	(0.015)***
BB (Pct. 75)	-0.418	-0.139
	(0.022)***	(0.016)***
Blind Math Score		1.010
		(0.034)***
Number of Observations	36044	36044
Number of Clusters	72	72
First-stage F Statistic		903.9

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. Column 2 follows the same specification from Table 2, column 4, except for the standard errors that here are calculated with school-level clusters.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B2) Estimated biases in the non-blind math scores toward classroom behavior while controlling for handwriting ability

	(1) IV	(2) IV	(3) IV
GB (Pct. 75)	0.096	0.091	0.087
	(0.018)***	(0.018)***	(0.018)***
BB (Pct. 75)	-0.130	-0.121	-0.120
	(0.017)***	(0.017)***	(0.017)***
Blind Math Score	1.026	1.011	1.029
	(0.032)***	(0.032)***	(0.032)***
Essay Scores	No	Yes	Yes
Instrumenting Essay Scores	-	No	Yes
Number of Observations	28338	28338	28338
Number of Clusters	12432	12432	12432
First-stage F Statistic	1168	1140	579.1

Note: This table reports student×exam-level IV (columns 1-3) regressions of teacher-assigned math scores on classroom behavior, in a subsample where essay scores are available. Column 1 follows the same specification from Table 2, column 4. Column 2 additionally controls for blind essay scores, and column 3 uses lagged essay scores as instrumental variable for the current ones.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B3) Estimated biases in the non-blind essay scores toward classroom behavior

	(1)	(2)
VARIABLES	OLS	IV
GB (Pct. 75)	0.406	0.210
	(0.024)***	(0.025)***
BB (Pct. 75)	-0.369	-0.166
	(0.023)***	(0.025)***
Blind Essay Scores		0.682
		(0.029)***
Number of Observations	15861	15861
Number of Clusters	6319	6319

Note: This table reports student×exam-level OLS (column 1) and IV (column 3) regressions of teacher-assigned essay scores on classroom behavior, for two different subsamples. One of them uses all the essay exams, and the other is restricted to essay scores that are high-stakes. $BB(Pct.75)$ and $GB(Pct.75)$ stand for binary variables that indicate whether students are at the top quartile of the essay behavior measures' distribution. All specifications include classroom fixed effects and exams fixed effects. Covariates additionally include indicators for age, gender, and ethnicity (Black, Indigenous, *Pardo*, Yellow, and White). Standard errors in parenthesis are robust and clustered at the student level.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B4) Estimated biases in the non-blind math scores toward classroom behavior while varying the timing differences between the exams

	(1)	(2)
VARIABLES	IV	IV
GB (Pct. 75)	0.122	0.126
	(0.031)***	(0.018)***
BB (Pct. 75)	-0.119	-0.146
	(0.028)***	(0.016)***
Blind Math Score	1.052	0.951
	(0.060)***	(0.039)***
Order of the exams	Blind first	Non-blind first
Number of Observations	11294	27833
Number of Clusters	4512	12469
First-stage F Statistic	374.1	793.5

Note: This table reports student×exam-level IV (columns 1-2) regressions of teacher-assigned math scores on classroom behavior, for two different subsamples according due to timing differences between the realization of the blind and non-blind exams. In column 1 (2), blindly-graded tests came first (after) the teacher-graded ones. All columns follow the same specification from Table 2, column 4. *** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B5) Estimated biases in the non-blind math scores toward classroom behavior while varying the timing differences between the exams (intensive margin)

	(1)	(2)
VARIABLES	IV	IV
GB (Pct. 75)	0.097	0.113
	(0.023)***	(0.017)***
BB (Pct. 75)	-0.141	-0.140
	(0.020)***	(0.015)***
Blind Math Score	0.941	1.012
	(0.046)***	(0.034)***
Timming difference	$<$ 3 weeks	$>$ 3 weeks
Number of Observations	18723	36273
Number of Clusters	14398	14838
First-stage F Statistic	598.3	1090

Note: This table reports student×exam-level OLS (column 1) and IV (columns 2-5) regressions of teacher-assigned math scores on classroom behavior, for two different subsamples according due to timing differences between the realization of the blind and non-blind exams. Column 1 (2) uses exams where the absolute difference varies less (more) than three weeks. All columns follow the same specification from Table 2, column 3. *** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B6) Estimated biases in the non-blind math scores toward classroom behavior while controlling ability

	(1)	(2)	(3)	(4)
VARIABLES	IV	IV	IV	IV
GB (Pct. 75)	0.108	0.112	0.108	0.111
	(0.027)***	(0.025)***	(0.026)***	(0.025)***
BB (Pct. 75)	-0.139	-0.142	-0.141	-0.143
	(0.025)***	(0.023)***	(0.024)***	(0.023)***
Blind Math Score	1.034	0.907	0.986	0.908
	(0.059)***	(0.076)***	(0.061)***	(0.076)***
Math Ability	No	Yes	No	Yes
Language Ability	No	No	Yes	Yes
Number of Observations	14140	14140	14140	14140
Number of Clusters	5565	5565	5565	5565
First-stage F Statistic	367.7	210.3	333.1	210.3

Note: This table reports student×exam-level IV (columns 1-4) regressions of teacher-assigned math scores on classroom behavior. Column 1 follows the same specification from Table 2, column 4; but restrict the sample to data available in both 2018 and 2019. Column 2 (3) additionally control for math (language) ability measured by the average scores in all blindly-graded math (language) exams from past year. Column 4 controls for both math and language ability.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B7) Testing whether the estimated biases in the non-blind scores toward classroom classroom behavior are statistically different under different ways of measuring behavior

	Good Behavior		Bad Behavior	
	Diff	P-value	Diff	P-value
Math 2018	0.03	0.32	0.00	0.91
All Subjects 2019	0.02	0.48	-0.02	0.42
All Subjects 2018	0.03	0.30	0.00	0.97
All Subjects - Math 2019	0.03	0.32	-0.01	0.57
All Subjects - Math 2018	0.04	0.26	0.03	0.35

Note: This table tests whether the estimate plotted in Figure 3, panel a (or b), under the label Math 2019 is statistically different from all the others presented within the same panel. More specifically, we test whether using the behavior reports from the math teacher from the past year, from all teachers from current or past year, or from all teachers from current or past year expect the math ones, lead to statistically different results than simply using the behavior assessments made by current math teachers. Diff is the observed difference between the coefficients. P-value is a boostrapped p-value from tests under the null that Diff equals zero.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

B OLS and IV Potential Biases

We consider a simple econometric model to analyze the bias of the OLS if the blind test score is measured with error, and of the IV estimator when the exogeneity assumption of the instrument does not hold. For simplicity, we assume that all variables have expected value equal to zero, we consider only the measure of good behavior, and we suppress the ijs sub-index. A simplified version of equation (1) is then given by

$$S^{NB} = \beta GB + \delta S^B + \varepsilon, \tag{3}$$

where $\varepsilon = \xi + r - \delta u + v$. We assume that $\mathbb{E}[GB\varepsilon] = 0$, but $\mathbb{E}[S^B\varepsilon]$ is potentially different from zero. Following the discussion from Section 3, we consider the case in which $r \approx 0$, so that $\mathbb{E}[S^B\varepsilon] \neq 0$ because of the measurement error u. We also assume that ξ is uncorrelated with GB and S^B.

Assuming that u is uncorrelated with all other variables in the model, we have that the OLS estimator is such that

$$\begin{bmatrix} \hat{\beta}^{ols} \\ \hat{\delta}^{ols} \end{bmatrix} \to_p \begin{bmatrix} \beta \\ \delta \end{bmatrix} + \frac{1}{\sigma_x^2\sigma_w^2 - (\sigma_{xw})^2} \begin{bmatrix} \sigma_{xw}\sigma_u^2\delta \\ -\sigma_x^2\sigma_u^2\delta \end{bmatrix},$$

where $\sigma_x^2 = var(GB)$, $\sigma_w^2 = var(S^B)$, $\sigma_u^2 = var(u)$, and $\sigma_{xw} = cov(GB, S^B)$. If we define the linear projection $GB = \gamma S^B + h$, then $\gamma = \frac{\sigma_{xw}}{\sigma_w^2}$. Therefore, $\sigma_x^2\sigma_w^2 - (\sigma_{xw})^2 = \sigma_x^2\sigma_w^2 \left[1 - \gamma^2\frac{\sigma_w^2}{\sigma_x^2}\right] > 0$, so the sign of the bias of the OLS estimator for β is determined by the signs of σ_{xw} and δ. Given model (1), we have that $\delta > 0$. Moreover, we can estimate σ_{xw} using the data, where we find $\hat{\sigma}_{xw} > 0$. Therefore, we should expect that $\hat{\beta}^{ols}$ is upward biased. The intuition is that the measurement error u implies that the estimator for δ will suffer from attenuation bias, which implies that it will not completely control for students' skills. If we consider instead our measure of bad behavior, then the correlation between BB and S^B is negative, which implies that the estimator associated with BB would be downward biased.

We consider next estimation of equation (3) using lagged blind test score LS^B as instrumental variable for S^B. This instrument clearly satisfies the relevance condition. If $\mathbb{E}[LS^B\varepsilon] = 0$, then the IV estimator would be consistent for β. We are worried, however, that the exogeneity condition for the instrument may not be valid. We assume that $\mathbb{E}[LS^B u] \approx 0$, which is a standard assumption in classical test theory and applied papers (e.g., Bond and Lang (2018)). Still, it may be that $\mathbb{E}[LS^B\xi] > 0$. For example, teachers may statistically discriminate students based on their past performance in blind scores or other correlated unobservable signals of scholastic ability. In this case, we have that the IV estimator will converge to

$$\begin{bmatrix} \hat{\beta}^{IV} \\ \hat{\delta}^{IV} \end{bmatrix} \to_p \begin{bmatrix} \beta \\ \delta \end{bmatrix} + \frac{1}{\sigma_x^2\sigma_{wz} - \sigma_{xw}\sigma_{xz}} \begin{bmatrix} -\sigma_{xw}\mathbb{E}[LS^B\varepsilon] \\ \sigma_x^2\mathbb{E}[LS^B\varepsilon], \end{bmatrix},$$

where $\sigma_{wz} = cov(LS^B, S^B)$ and $\sigma_{xz} = cov(LS^B, GB)$. Note that $\sigma_x^2\sigma_{wz} - \sigma_{xw}\sigma_{xz} = cov(e_1, e_2)$, where e_1 is the population error in the linear projection of GB on S^B, and e_2 is the population error in the linear projection of GB on LS^B. If we consider the residuals from a regression of GB on S^B and the residuals from a regression of GB on LS^B, then the correlation between these two residuals is positive, which provides evidence that $\sigma_x^2\sigma_{wz} - \sigma_{xw}\sigma_{xz}$ is positive. Given that $\sigma_{xw} > 0$ when we consider a measure of good behavior, if we have $\mathbb{E}[LS^B u] \approx 0$ and $\mathbb{E}[LS^B \xi] > 0$, then $\hat{\beta}^{IV}$ would be downward biased. Likewise, if we consider a measure of bad behavior, then the estimator associated with this variable would be upward biased.

Combining these results, we have that the discrimination parameters are bounded by the OLS and the IV estimators.

C Racial and Gender Discrimination

Table C1 presents the estimated bias in the non-blind math scores toward black pupils. Blacks' average non-blind scores are 0.23 SD below than whites' (column (1)). This gap falls drastically (-0.077, s.e. 0.042) when we adjust for student proficiency captured by blind scores. Our estimates have relatively higher magnitude in comparison to Botelho et al. (2015) and are very similar to Hanna and Linden (2012). The first also find grading biases that harm black pupils in a Brazilian sample with a high share of black students by contrasting teacher-assigned end-of-year grades and state proficiency scores. The second uses data from India and, in a correspondence study design, find grading biases toward students' caste.

In Table C2, we present the estimated grading bias toward gender. Boys have advantages of 0.03 SD over girls in math test scores. This could indicate some favoritism toward boys. However, by controlling for non-blind math grades we find evidence that boys' math proficiency is under-assessed by teachers. The grading bias is equivalent to a taxation of 0.08 SD in blind scores. Our main estimate drops significantly when we control for student in-class behaviors, reflecting the fact that boys have worse classroom behavior, but remains statistically different from zero. These findings are in line with previous studies using a similar research design (Lavy, 2008; Falch and Naper, 2013).

Taken together, our results indicate that despite suggestive evidence of grading biases against boys and black students, results are much lower in comparison to biases toward behavior. Additionally, we do not find any heterogeneous effects across race and gender. Finally, Figure C.1 shows we also find grading biases against boys and black students in other subjects.

Table (C1) Estimated biases in the non-blind math scores against black pupils

	(1)	(2)	(3)	(4)
VARIABLES	OLS	IV	IV	IV
Black	-0.235	-0.077	-0.075	-0.089
	(0.048)***	(0.042)*	(0.041)*	(0.049)*
GB (Pct. 75)			0.112	0.107
			(0.017)***	(0.020)***
BB (Pct. 75)			-0.140	-0.135
			(0.015)***	(0.018)***
Black x GB				0.086
				(0.095)
Black x BB				-0.009
				(0.098)
Blind Math Score		1.045	1.010	1.010
		(0.033)***	(0.034)***	(0.034)***
Number of Observations	36044	36044	36044	36044
Number of Clusters	14692	14692	14692	14692
First-stage F Statistic		1162	1098	1097

Note: This table reports student×exam-level OLS (column 1) and IV (columns 2–4) regressions of teacher-assigned math scores on ethnicity. Black stands for a binary variable that indicates whether student is black. We also control for other ethnic indicators. The omitted category is white. $BB(Pct.75)$ and $GB(Pct.75)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In the IV estimates, lagged blind math scores are used as instrumental variable for the current math scores. Columns 2-4 also controls for past blind scores of language, science, and humanities. All specifications include classroom fixed effects and exams fixed effects. Standard errors in parenthesis are robust and clustered at the student level.

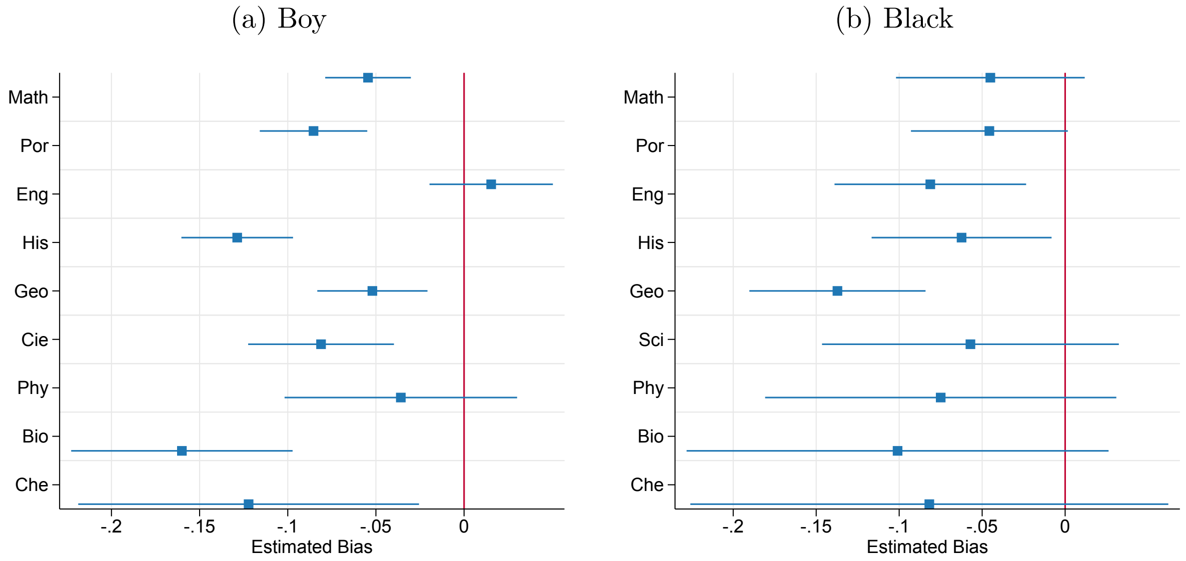
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (C2) Estimated biases in the non-blind math scores against boys

	(1)	(2)	(3)	(4)
VARIABLES	OLS	IV	IV	IV
Boy	0.033	-0.080	-0.054	-0.052
	(0.015)**	(0.012)***	(0.012)***	(0.016)***
GB (Pct. 75)			0.112	0.121
			(0.017)***	(0.020)***
BB (Pct. 75)			-0.139	-0.145
			(0.015)***	(0.022)***
Boy x GB				-0.023
				(0.026)
Boy x BB				0.010
				(0.028)
Blind Math Score		1.045	1.010	1.010
		(0.033)***	(0.034)***	(0.034)***
Number of Observations	36044	36044	36044	36044
Number of Clusters	14692	14692	14692	14692
First-stage F Statistic		1163	1099	1098

Note: This table reports student×exam-level OLS (columns 2–4) and IV (column 2) regressions of teacher-assigned math scores on gender. $BB(Pct.75)$ and $GB(Pct.75)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In the IV estimates, lagged blind math scores are used as instrumental variable for the current math scores. Columns 2-4 also control for past blind scores of language, science, and humanities. All specifications include classroom fixed effects and exams fixed effects. Standard errors in parenthesis are robust and clustered at the student level. *** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Figure (C.1) Estimated biases in the non-blind scores from Mathematics, Portuguese, English, History, Geography, Science, Physics, Biology, and Chemistry



(a) Boy

(b) Black

Note: This figure plots 95% confidence intervals computed with student-level cluster and point estimates from student×exam-level IV regressions of teacher-assigned scores from several subjects on indicators for boys and black students, and proficiency in the material covered by the examination (measured by the blind scores, which we instrument using lagged blind scores). When the dependent variable is a math test score, we measure proficiency using the blind math scores, and additionally control for past performance in blindly-graded science and language exams. When the dependent variable is a Portuguese or English test score, we measure proficiency using blind language score, and additionally control for past performance in blindly-graded math and science exams. When the dependent variable is a teacher-assigned score in science, chemistry, physics, or biology, we measure proficiency using blind science score, and additionally control for past performance in blindly-graded math and language exams. All specifications additionally include controls for age, classroom behavior, other ethnic indicators (the omitted category is white) and fixed effects for classroom and exam. Students from grades 6 through 9 do not have classes in chemistry, physics, and biology; only a science class. The opposite is true for students from grades 10 through 11.

D Further Exploiting Behavior Data

D.1 Alternative moments of the distribution and continuous measures

The results presented in this paper are based on a discretization of the behavior measures. One might be worried whether our evidence depends strongly on this transformation. Table D1 presents similar results when we use other moments of the distribution. Moreover, Tables D2 and D3 depict a similar pattern when we use the continuous measures. The first presents the results when we use compute the measures using the behavior assessments from the previous cycle only. The second uses the accumulated reports. The share of the association between behaviors and test scores explained by grading biases remains very similar to our main results: 20% for the good behavior and 40% for the bad behavior. Finally, we also obtain similar results when using the overall number of good and bad assessments as regressors (see Table D4).

Table (D1) Estimated biases in the non-blind math scores toward classroom behavior

	(1)	(2)	(3)	(4)
VARIABLES	IV	IV	IV	IV
GB (Pct. 50)	0.598	0.147		
	(0.017)***	(0.018)***		
BB (Pct. 50)	-0.348	-0.107		
	(0.016)***	(0.016)***		
GB (Pct. 90)			0.766	0.099
			(0.025)***	(0.026)***
BB (Pct. 90)			-0.545	-0.187
			(0.024)***	(0.025)***
Blind Math Score		1.029		1.057
		(0.038)***		(0.039)***
Number of Observations	25085	25085	25085	25085
Number of Clusters	14686	14686	14686	14686
First-stage F Statistic		498.8		505.8

Note: This table reports student×exam-level OLS (columns 1 and 3) and IV (columns 2 and 4) regressions of teacher-assigned math scores on classroom behavior. $BB(Pct.50)$ and $GB(Pct.50)$ stand for binary variables that indicate whether students are above the median of the math behavior measures' distribution. $BB(Pct.90)$ and $GB(Pct.90)$ stand for binary variables that indicate whether students are above the 90th percentile of the math behavior measures' distribution. Columns 2 and 4 follows the same specification from Table 2, column 4.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (D2) Estimated biases in the non-blind math scores toward classroom behavior – using the continuous behavior measures (cycle-specific lagged reports)

	(1)	(2)
VARIABLES	OLS	IV
GB	0.895	0.205
	(0.026)***	(0.028)***
BB	-0.335	-0.131
	(0.026)***	(0.028)***
Blind Math Score		0.965
		(0.041)***
Number of Observations	21804	21804
Number of Clusters	12641	12641
First-stage F Statistic		734.7

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. *BB* and *GB* stand for the math behavior measures, computed using the behavior assessments from the previous cycle. Column 2 follows the same specification from Table 2, column 4.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (D3) Estimated biases in the non-blind math scores toward classroom behavior – using the continuous behavior measures (accumulated reports)

	(1)	(2)
VARIABLES	OLS	IV
GB	1.143	0.243
	(0.027)***	(0.027)***
BB	-0.426	-0.178
	(0.026)***	(0.024)***
Blind Math Score		0.993
		(0.034)***
Number of Observations	36044	36044
Number of Clusters	14692	14692
First-stage F Statistic		1079

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. *BB* and *GB* stand for the math behavior measures, computed using all the behavior assessments that preceded the math exam. Column 2 follows the same specification from Table 2, column 4.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (D4) Estimated biases in the non-blind math scores toward classroom behavior – using the number of behavior assessments

	(1)	(2)
VARIABLES	OLS	IV
Ln GB Reports	0.345	0.090
	(0.009)***	(0.009)***
Ln BB Reports	-0.211	-0.057
	(0.008)***	(0.008)***
Blind Math Score		1.018
		(0.034)***
Number of Observations	38461	38461
Number of Clusters	15553	15553
First-stage F Statistic		1102

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. Ln BB reports and Ln GB reports stand for the natural logarithm of the number of bad and good behavior assessments received by math teachers plus 1. Column 2 follows the same specification from Table 2, column 4. *** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

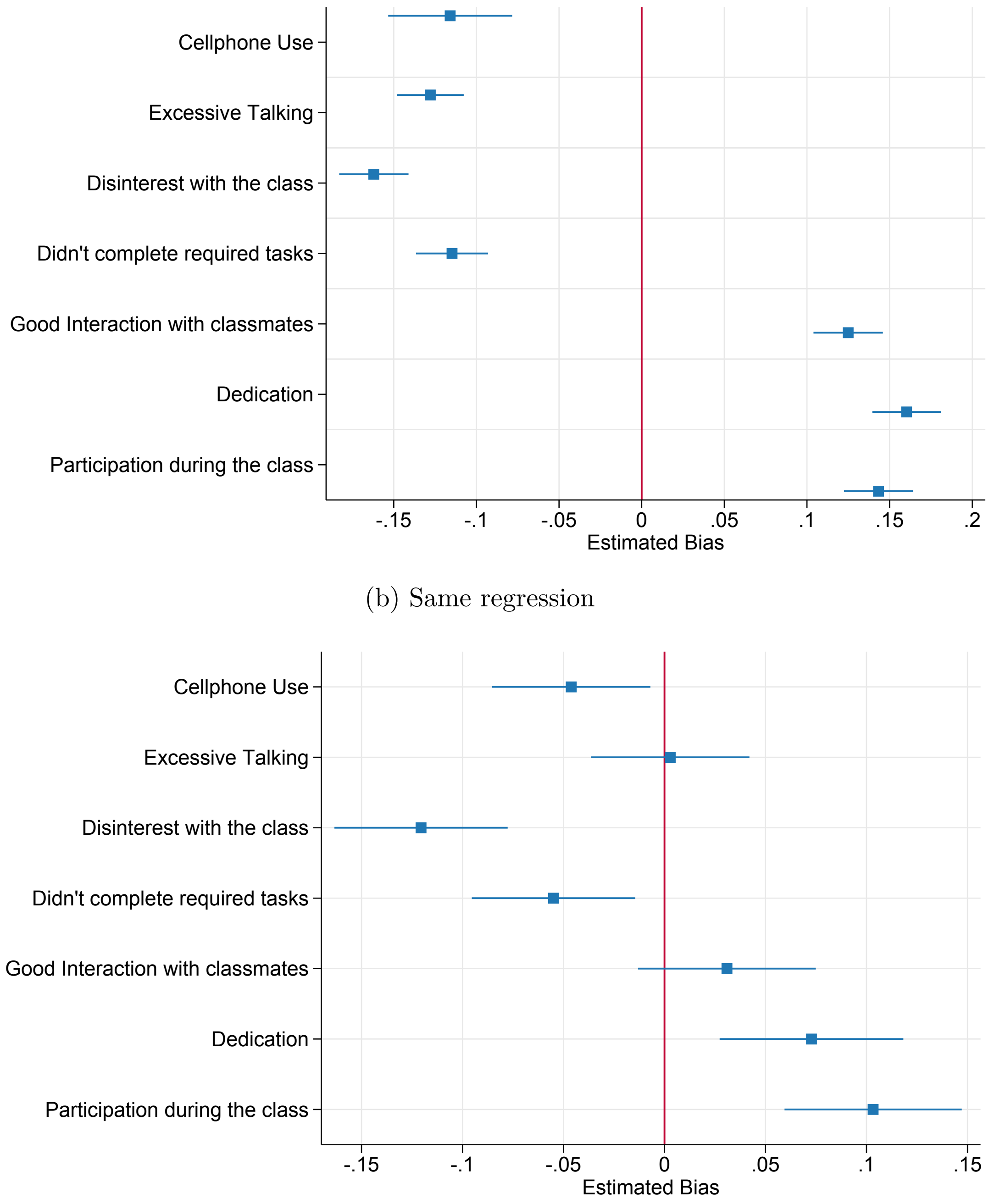
D.2 What are the behaviors driving the results?

Here we estimate heterogeneous effects for each of the behaviors. To do so, we calculate dis-aggregated behavior measures. Take the behavior report "Dedication" as an example. Let d_{ijs} indicate the number of assessments i received under this category by a subject s teacher until exam j. The measure $Dedication_{ijs}$ is then defined as:

$$Dedication_{ijs} := \frac{d_{ijs}}{\max\{d_{hjs} : h \in \mathcal{C}(i)\}}.$$

Figure D.1 presents our main estimates. In panel (a), we estimate the grading biases toward each of the behaviors separately. Overall, the point estimates are very similar, indicating that they are all capturing correlated biases. In panel (b), we estimate the effects using the same regression model. The negative discrimination is driven by the disinterest of students during the class (-0.12, s.e. 0.026) and is followed by 'Did not complete the required tasks' (-0.054, s.e. 0.024), and cellphone use (-0.025, s.e. 0.014). The positive discrimination is driven by participation during the class (0.1, s.e. 0.026), and is followed by dedication (0.07, s.e. 0.026), and good interaction with classmates (0.03, s.e. 0.025).

Figure (D.1) Estimated biases toward each behavior



(a) Different regressions

(b) Same regression

Note: These figures plots student×exam-level IV regressions of teacher-assigned math scores on measures for each classroom behavior. Panel (a) plots point estimates from different IV regressions on each behavior. Panel (b) plots point estimates from an IV regression on all the behaviors. Both also plots 95% confidence intervals. All regressions follow the same specification from Table 2, column 4.

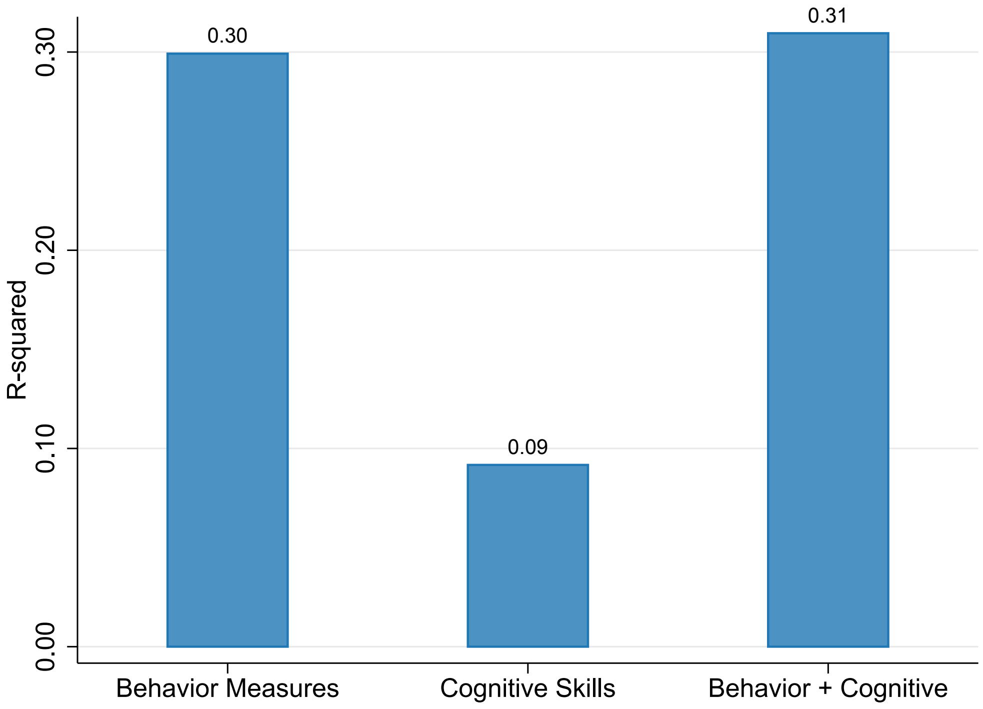
D.3 Socio-emotional assessments

For a subset of our schools, students take a regular course designed to improve six socio-emotional abilities: grit, creativity, cooperation, communication, pro-activity, and critical thinking. During the course, pupils learn about daily habits that should lead them to better understand a particular socio-emotional ability, reflect on it, and finally act according to it. Classes are guided by games, projects, and audiovisual content, created by our partner specifically for this course. The bulk of the material is a teen series that brings important themes to classroom discussions. In each class, students discuss socio-emotional skills through the series' characters' habits and challenges.

In this course, there are no formal examinations. Still, students also earn scores. They are evaluated through projects within their material and in-class activities. We standardize the assessments from each cycle and use it as a measure of non-cognitive abilities. Figure D.2 plots how much of the average socio-emotional score's variability is explained by a linear model on the behavior measures, cognitive skills measured by the average score on blindly-graded achievement tests, and both the behavioral and cognitive components. The behavior measures explain a high proportion of the socio-emotional score's variance: 30%. The predictive power of the cognitive skills is smaller: 9%. Moreover, after controlling for student behaviors, the effect of cognitive abilities becomes negligible. Hence, the socio-emotional score reflects students' non-cognitive characteristics highly correlated with in-class behaviors.

We then follow the empirical strategy outlined in Section 3, but use the socio-emotional factor to capture student behaviors. To compute the socio-emotional score, we use only the assessments that preceded each teacher-graded math examination. Table D5 presents the results. One SD increase in the socio-emotional score is associated with increased teacher-assigned test scores of 0.25 SD (columns 1). When we control for the cognitive skills measured in those tests (column 2), the socio-emotional effect is significantly reduced. Further controlling for student demographics (column 3) and other cognitive abilities measured by performance in other blind exams (column 4) barely change the results. Overall, moving the socio-emotional score by one SD produce a test score bonus of 0.084 SD when achievement exams are graded by teachers. This represents 35% of the unconditional correlation. Table D6 shows the results do not change if we measure socio-emotional abilities using the scores students earned in the previous year by a different set of teachers. Taken together, these results reinforce the point that teachers factor in non-achievement factors when assessing their students' cognitive abilities in achievement examinations.

Figure (D.2) Predictive power of the behavior measures and cognitive skills on the socio-emotional score



Note: This figure plots the adjusted R-squared from different linear regressions where the dependent variable is the socio-emotional score. In the first model, the explanatory variables are the behavior measures. In the second, cognitive skills measured by the average of all the blindly-assigned test scores. In the third one, we add both the behavior measures and the cognitive skills.

Table (D5) Estimated biases in the non-blind math scores toward socio-emotional abilities

	(1)	(2)	(3)	(4)
VARIABLES	OLS	IV	IV	IV
Socio-emotional score	0.247	0.095	0.086	0.084
	(0.016)***	(0.013)***	(0.014)***	(0.014)***
Blind Math Score		1.023	1.023	1.015
		(0.023)***	(0.023)***	(0.057)***
Student Demographics	No	No	Yes	Yes
Other Scores	No	No	No	Yes
Number of Observations	9902	9902	9902	9902
Number of Clusters	3411	3411	3411	3411
First-stage F Statistic		2062	1956	381

Note: This table reports student×exam-level OLS (column 1) and IV (columns 2-4) regressions of teacher-assigned math scores on socio-emotional skills. Socio-emotional score stands for the standardized assessments students received in coursers designed to improve their non-cognitive skills in a period that preceded the exam. In the IV estimates, lagged blind math scores are used as instrumental variable for the current math scores. All specifications include classroom fixed effects and exams fixed effects. Other scores include the cumulative average performance in science and humanities, and current performance in language. High-order polynomials for scores include a third order polynomial for blind math scores, and an interaction term between math and language scores. Controls include indicators for age, gender, and ethnicity (Black, Indigenous, *Pardo*, Yellow, and White). We also include a dummy for students with missing data on ethnicity. Standard errors in parenthesis are robust and clustered at the student level.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (D6) Estimated biases in the non-blind math scores toward socio-emotional abilities – using the assessments received in the current and past year

	(1)	(2)
VARIABLES	IV	IV
Socio-emotional score (19)	0.095	
	(0.024)***	
Socio-emotional score (18)		0.095
		(0.029)***
Blind Math Score	1.117	1.120
	(0.112)***	(0.112)***
Number of Observations	3184	3184
Number of Clusters	1075	1075
First-stage F Statistic	112.8	112.5

Note: This table reports student×exam-level IV (columns 1–2) regressions of teacher-assigned math scores on socio-emotional skills. Socio-emotional score (19) is computed as in Table D5. Socio-emotional score (18) is computed using the assessments students received in the socio-emotional course taken in the past year. Columns 1–2 follow the same specification from Table D5, column 4.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.